When Naturalisms Collide: Neural Representations as Scientific Posits and Metaphysical Entities

***Article Forthcoming in Philosophy and the Mind Sciences*.**

Eric Hochstein

Abstract: Can neural representations be naturalized? Current debates surrounding the naturalization of representations in neuroscience typically characterize this project in terms of one of three options: methodological naturalism, ontological naturalism, or the belief that both types of naturalism provide support for, and constraints on, each other to drive inquiry. In this paper, I argue that all three of these options are problematic. The two projects of naturalism cannot be pulled apart from one another, nor can one act as effective support/constraint on the other. The relationship between these projects of naturalism is far more complex, nuanced, and interdependent than is typically thought. I highlight how this influences current debates regarding the nature of representation in neuroscientific practice.

Keywords: Naturalism, Ontological Naturalism, Methodological Naturalism, Representation, Intentionality.

1. Introduction

The quest to naturalize intentionality is hardly new in philosophy (e.g. Dretske, 1981, 1988; Stalnaker, 1987; Fodor, 1987; Millikan, 1984, 1989). However, recent debates within philosophy of neuroscience and cognitive science have rekindled interest in such a mission, albeit in a slightly different form (e.g. Thomson & Piccinini, 2018; Williams, 2018; Shea, 2018; Morgan & Piccinini, 2018; Segundo-Ortin & Hutto, 2021; Poldrack, 2021; Favela, 2021; Taylor, 2022; Nanay, 2022; Favela & Machery, 2023; Ward, 2023). These disputes center around the question of whether we should understand various neurological and cognitive processes as representational in nature or not. To show that the brain is truly forming and manipulating mental representations to solve problems and navigate the world, one must show that representation is a natural phenomenon. Otherwise, it calls into question whether it is a legitimate posit for neuroscientists, cognitive scientists, and philosophers to invoke when studying the brain and understanding human cognition. As Samuel Taylor summarizes it:

> Representationalists and anti-representationalists disagree about whether a naturalization of mental content is possible and, hence, whether positing mental representations in cognitive science is justified. (2022, p. 518)

This problem is essentially the problem of intentionality in new garb.[1] As Robert Stalnaker puts it, “the Problem of Intentionality is a problem about the nature of representation” (1987, p. 6).

The project of naturalizing mental representation or intentionality is often linked with the project of realism. To show that mental representation is a natural phenomenon is to show that representations deserve a place within our ontology. The method by which we can show that such a phenomenon is natural is often characterized in one of two ways. As Taylor puts it:

> Here, I think we have only two options: that a successful naturalisation of [representational] content would give an account of content in terms of physical and/or spatiotemporal entities (call this ontological naturalisation); or that a successful naturalisation of content would give an account of content in terms of the entities that play a role in (cognitive) scientific explanations (call this methodological naturalisation). (2022, p. 523)

Before proceeding further, it is worth making a note of clarification here. Taylor’s phrasing can be somewhat unclear. *Both* accounts of naturalism are interested in what sorts of entities should be included within our ontology, and so calling only one account “ontological naturalism” is misleading. What differentiates them is that methodological naturalists do not commit themselves to any particular set of metaphysical properties that an entity must have in order to be real, instead relying on the

[1] Original disputes about naturalizing intentionality were frequently tied up with the existence of *propositional attitudes*. More recent debates in cognitive science do not define representations in terms of propositional attitudes. However, the issue of naturalizing representational content remains the same.

success of the scientific theories which invoke such entities to determine what is real. Meanwhile, ontological naturalists have additional metaphysical conditions for what sorts of entities we should be permitted to consider real (i.e. they must be physical, have causal powers, be spatio-temporally located, etc). To avoid confusion, I will follow Penelope Maddy (2011) in labeling this second form of realism "robust realism", to highlight that not any kind of metaphysical entity or process invoked by our best scientific theories count as natural for the ontological naturalist, only those that meet certain metaphysical standards.[2]

These two projects of naturalization are often taken to have different aims and goals, one focused on the explanatory value of representations as invoked within successful neuroscientific theories (methodological naturalism), the other on whether "representation" denotes a set of robust entities and processes located in the human brain with intentional powers (ontological naturalism). For this reason, it is often argued that these projects should not be conflated, and that to engage in one project of naturalization is not necessarily to engage in the other (e.g. Shapiro, 1997; Collins, 2015; Mendelovici, 2018; Tayler, 2022). As a result, the two projects of naturalism are typically thought to relate in one of two ways: (1) the two projects are largely independent of each other, focusing on distinct senses of "natural". Being natural in one sense (e.g. methodological) is not the same as being natural in the other (e.g. ontological). Thus, we can engage in one project of naturalism while ignoring the other. Or (2) The two projects provide essential support for, and constraints, on one another in developing a scientific understanding of what neural representations are and how they come about.

[2] Maddy contrasts "robust realism" with what she calls "thin realism", which she argues is justified by the methodological naturalist project. Using the existence of mathematical sets as an example, she describes the difference between them as follows:

> So the fundamental diagnostic is this: the Robust Realist requires a non-trivial account of the reliability of set-theoretic methods, an account that goes beyond what set theory tells us; for the Thin Realist, set theory itself gives the whole story; the reliability of its methods is a plain fact about what sets are. (2011, p. 63)

In this paper I argue that contrary to such views, these two projects are not truly distinct from one another (and so we cannot ignore one project to engage in the other), but nor can we use one project to provide effective constraint or support for the other. To engage in one project of naturalism already requires liberally taking from the other project at every step of inquiry. Each project is so baked into each other, that attempts to use one project to support or constrain the other end up being circular as each project already presupposes views from the other. This fact forces us to reconsider our understanding of what it means to justify neural representations as both an entity invoked in our best scientific theories, and as a robust metaphysical phenomenon.

In order to make this argument, I start in section 2 by highlighting the two ways in which "naturalization" is most commonly understood within contemporary debates regarding neural representations, and why there seem to be compelling reasons to think of these projects as distinct from one another. In section 3, I highlight how the two seemingly distinct projects necessarily and unavoidably distort, modify, and structure, each other in subtle and overt ways, and why we cannot bracket one type of naturalism to only focus on the other. In section 4, I turn to the idea that these two projects can work in concert to provide supporting evidence for, and constraints on, one another in our understanding of neural representations. I argue that this view is unattainable since each project is so heavily structured by, and intertwined with, the other that their results are not sufficiently independent of each other to provide the necessary support or constraint. I conclude by talking about the implications of this for current disputes regarding the naturalization of representations in neuroscience and cognitive science.

2. The two projects of naturalism

Naturalism has a long history in philosophy, but there is substantial variation in what people mean by "naturalism", and what implications it has for science and metaphysics (see: Flanagan, 2006). My intention here is not to provide a survey of the different senses of naturalism that have been proposed over the years, but instead to highlight two distinct senses of naturalism that have tended to dominate disputes about representation in neuroscience and cognitive science. To begin, let's consider each of the two projects of naturalism in isolation, and why some have proposed that we should think of these projects as distinct from one another.

2.1 Methodological Naturalism

The first type of naturalism, often called "Methodological Naturalism" (MN) can be understood roughly as follows:

> The pursuit of science on this conception is to fill out the model of the 'universe' incrementally, but at no point are we to assume on metaphysical grounds that the 'essence' or 'cause' of a given phenomenon must give way to some list of fundamental properties. Thus it was that Newton could boldly present universal gravitation as, effectively, an all-pervading 'action at a distance' force. Newton found such a force highly problematic, but he certainly didn't think of it as 'occult' as did his European critics; on the contrary, on the new conception of science, one should expect such abstractions and await their eventual integration into some as yet perhaps undreamt of scheme. [...] Such a general take on Newton's achievement, whereby theories are no longer constrained to make sense according to prevailing metaphysics and/or common sense, but only in terms of their consistency, simplicity and economy, and prediction of relevant data, has come to dominate scientific thinking. (Collins, 2015, p. 90-91)

Put another way, if one thinks that science ought to be our guide for determining what exists, then we must not make the mistake of thinking we can somehow step *outside* the methods of science in order to pass judgement on the sorts of entities that really exist, and thus which ones science *ought* to invoke. If science is to be our guide for what is real, then we must work entirely *within* the confines of scientific practice. In this way, the successes of our scientific theories and practices must be our primary guide for justifying which entities are real and which are not, regardless of whether they turn out to be physical, causally efficacious, spatiotemporally localizable, or something else altogether. Maddy (2011), for instance, makes this argument regarding the existence of mathematical entities like sets. She tells us that:

> Since set theory tells us nothing about sets being dependent on us as subjects, or enjoying location in space or time, or participating in causal interactions, it follows that sets are abstract in the familiar ways. [...] In addition to the familiar concrete objects [scientists have] been studying so far, there are also objective, non-spatiotemporal, acausal sets. (2011, p. 62)

Put simply, methodological naturalism is the view that the theoretical entities which should be granted a place in our ontology be determined by those that play a necessary role in our current best explanatory theories in science, and not pre-judged as natural based on metaphysical considerations of which entities are to be deemed respectable. This is often what philosophers who endorse MN have in mind in the context of neural or cognitive representations. Lawrence Shapiro, for example, claims that "the question whether intentionality is natural reduces to the question whether cognitive science is in fact a [successful] science" (1997, p. 320). Likewise, Andrew Richmond (2024) argues that:

> Which neural activity represents what? Which things should we understand as representations? Does FFA represent faces or some broader domain? The answer is that we should understand brain activities/structures in representational terms wherever it is helpful to apply the

> explanatory strategies that come with representational notions, and we should understand the brain as representing whatever domain provides the best models for it, given our interests and questions. (p. 16-17).

Instead of first determining what sorts of entities are metaphysically respectable, and then using this as a guide for what theoretical entities our scientific theories should invoke, MN argues that we first look at which scientific theories are successful, which in turn justifies the sorts of entities we should consider natural and thus real. Abstracta, purely mathematical objects, and the like would all be worthy of inclusion in our ontology just so long as they are vindicated by the success of the theories which invoke them (e.g. Dennett, 1991; Maddy, 2011). Similarly, one need not have any sort of metaphysical account of how the entities invoked by successful scientific theories arise from more primitive metaphysical entities/processes. Since this assumes that we first must determine these metaphysical criteria prior to whether science should invoke them, but this again gets the story backwards. As Shapiro puts it, "rather than decree that whatever is made of the right ingredients is natural, that is, we should allow the natural to be defined by the methods of science" (1997, p. 318). This means we need not have any sort of account of how representations or intentionality must be instantiated in the brain, or whether they can be understood in terms of more fundamental non-intentional entities or relations.

The real issues surrounding whether we can naturalize representation under this account hinge on whether representations truly are invoked in our best explanatory theories in neuroscience and cognitive science (e.g. Eliasmith, 2012; Clark, 2015; Weiskopf, 2017; Williams, 2018; Favela, 2021). Note that such debates focus on the explanatory and methodological value that theories which invoke representations have to neuroscientific practice.

2.2 Ontological Naturalism

Such debates can be contrasted with a different project of naturalism: Ontological Naturalism (ON). This second project is not concerned with whether successful neuroscientific or cognitive explanations happen to posit “representations” as entities in their theories, but with the question of whether such posits denote causally efficacious physical states/processes operating in the brain with intentional powers. Those involved with this project attempt to understand how representations (understood in this way) are possible in a world governed by physical objects and laws of nature that have no content or directedness. As Mendelovici & Bourget put it, this project is concerned with determining “what intentionality really is using only naturalistically acceptable entities, such as causal relations, evolutionary histories, and the like” (2014, p. 326). Shapiro describes this kind of naturalism as “Lego Naturalism”, since we naturalize intentionality by showing how it can be constructed or built out of more fundamental and uncontroversially natural units. As he puts it,

> Lego naturalists see the project of naturalizing intentionality as requiring that one show how to build it from the right kinds of pieces, where the right kinds of pieces are those that our nonintentional, natural, sciences describe. The surest way to naturalize the mind, the Lego naturalist believes, is to show how to build it from pieces that our nonintentional sciences have already certified as natural. (1997, p. 309)

This ontological project of naturalism is importantly different from the methodological project, both in terms of their goals and their methods. While MN argues that we can naturalize representations without saying anything about how such representations are instantiated by the brain, or what the exact metaphysical relationship is between representational content and the world, ON argues that it is precisely these things that determines whether representations are natural and thus real.

Attempts to naturalize representation within the context of this second ontological project have garnered a great deal of attention in traditional analytic philosophy of mind. Fred Dretske, for instance, famously attempts to naturalize representational content by showing how it can be constructed out of a more fundamental non-intentional sense of information, one rooted in information theory (Dretske, 1981, 1988). Meanwhile, Ruth Millikan (1984, 1989) attempts to naturalize intentionality and mental representation by appealing to the evolutionary history of the organism, and how this shaped the neurological mechanisms in question to carry out particular functions (and thus carry particular content). More recent attempts to ontologically naturalize representation in the context of neuroscience and cognitive science often build on these early informational and teleological accounts of representation (e.g. Cao, 2012; Gładziejewski & Miłkowski, 2017; Morgan & Piccinini, 2018; Thomson & Piccinini, 2018).

Meanwhile, those who argue against the existence of neural representations from within the context of this project argue that no acceptable or satisfactory metaphysical story of how such representational content is actually realized by the natural world has ever truly been provided (e.g. Segundo-Ortin & Hutto, 2021). As such, our best metaphysical account of the world tells us that representations do not exist, nor are they necessary to produce the mental and physical activities we engage in (e.g. Chemero, 2009; Hutto and Myin, 2013, 2014; Stewart, Gapenne & Di Paolo, 2010).

Note that these two projects are guided by different methods. For the methodological naturalist, it is the successes of our scientific theories that act as the primary method for determining what entities ought to be considered natural and thus real (regardless of whether they are abstract objects, physical processes, or something else altogether). For the ontological naturalist, on the other hand, this method is insufficient to justify the natural status of entities. This is because our successful neurological theories may be instrumentally valuable but false, or may demonstrate that representations are at best useful fictions. The methods of ON require that we show how representations can be constructed from the physical and causal processes in the world that science has already vindicated (a method that MN considers neither necessary nor sufficient for naturalization).

2.3 Keeping our naturalisms distinct

For our purposes, it is the distinction between these two projects of naturalization that is our focus. It is frequently argued that we should not confuse MN with ON when accounting for representations. As Angela Mendelovici puts it:

> There are interesting questions in the philosophy of science surrounding the notions of representation operative in various disciplines and research programs. What are these notions of representation? What roles do they play? Do different research programs use the same notion of representation? Some philosophers explicitly claim to be trying to answer these types of questions and not the types of [metaphysical] questions I'm concerned with. […] One prima facie reason to think this might be the case is that it makes sense to ascribe at least some of the kinds of representational states operative in the mind-brain sciences to artifacts that we might not really believe to have genuine intentional powers, such as calculators and computers. […] It could also turn out that the two ways of defining "intentionality" do not pick out the same thing because the definition based on the mind-brain sciences does not pick out anything at all. Perhaps the best understanding of talk of representation in the mind-brain sciences takes representational notions to be merely a dispensable fiction. (2018, p. 10-11)

The general idea is that there are different questions at play here. One is whether a scientific theory which invokes representations in neuroscience and cognitive science is good or successful (and thus whether we are scientifically licensed to take its posits seriously). The second is whether these posits denote robust physical states and processes in the brain with intentional powers. These two questions need to be demarcated.

While ontological naturalists might be happy to grant that some representations exist as abstract entities, and so are real in *that* sense, they are simply not interested in that sense of “real” when they seek to naturalize intentionality or representation in neuroscience or cognitive science. They seek something more robust, more “metaphysically committed” as Gładziejewski & Miłkowski (2017, p. 39) put it. Likewise, the methodological naturalist might be happy to grant that all kinds of physical causal entities exist, but deny that for representations in neuroscience to count as natural, and thus real, they *must* be understood in terms of robust spatiotemporal entities with intentional powers. In this regard, each project seems to be interested in a different sense of “natural” and “real”, and so each can go about its business without concern for the other.

For the ontological naturalist, the role that idealization and abstraction plays in scientific theorizing means that the sorts of entities posited by successful science are not always a good guide for determining what is robustly real. For instance, the ideal gas law is commonly used to predict the behaviour of real-world gasses by describing them as “ideal gasses”, which we know do not exist as robust entities in nature. Meanwhile evolutionary biology frequently employs infinite population sizes within their models and theories, when we know that no population in nature is ever infinite. Scientists and philosophers can debate the explanatory value, and potential indispensability, that “ideal gasses” or “infinite population sizes” may have to scientific theories and models, while also acknowledging that the outcome of these debates have no bearing on debates about whether we should believe that ideal gasses and infinite population sizes exist as robust entities out in the world. Rosenberg (2015), for instance, notes that there are many theorists who treat mental representation and other intentional concepts in exactly this way. He says:

> The eliminativist acknowledges the neo-Behaviorist's observation that intentionality is a predictively useful tool, a stance we employ. It's not just useful, it's indispensable for creatures like us. […] It's worth sketching to show why a stance can be indispensible even when its attributions are false. (p. 542)

Similarly, Adrian Downey argues that "representation can play an epistemologically indispensable role within predictive processing explanations [of cognition] without thereby requiring that representation metaphysically exists" (2018, p. 5115). This implies that whether representations are natural in the methodological sense is not the same as whether they are natural in the ontological sense. Moreover, the conditions by which we judge the success of one project of naturalism are not the same as the conditions by which we judge the success of the other. For instance, Weiskopf (2017) argues that we evaluate the explanatory success of scientific models and theories that invoke representations without making any sort of claims as to how such representations are constituted by the physical processes and structures of the brain. As he puts it, "cognitive models are capable of giving explanations of their target phenomena that answer to all of the relevant epistemic norms and standards, and they achieve this without making essential reference to the details of those models' neural implementation" (2017). To this end, MN certainly seems to come apart in a rather straightforward way from ON.

The idea that these are two distinct naturalizing projects with distinct aims and goals that should not be confused for one another has been presupposed, in one form or another, by philosophers for decades. Shaprio (1997), for instance, tells us that "[ontological] natural kinds cross-classify methodologically natural kinds" (p. 319-320). Taylor similarly notes that because there are entities that "could play a role in methodological, but not ontological, naturalisation demonstrates that the two kinds of naturalisation can and should be differentiated" (2022, p. 523). Likewise, Thomson & Piccinini (2018) draw a distinction between neural representations that exist as "mere theoretical posits" (p. 223) in successful neuroscientific theories, with neural representations that are "as real as neurons, action potentials, or any other well-established entities in our ontology" (p. 191). John Searle (1992) demarcates this distinction in terms of "literal" descriptions of intentionality or representation, with "as-if" or "metaphorical" descriptions of intentionality. In his words:

> [There is a] distinction between the sort of facts corresponding to ascriptions of intrinsic intentionality and those corresponding to as-if metaphorical ascriptions of intentionality. There is nothing harmful, misleading, or philosophically mistaken about as-if metaphorical ascriptions. The only mistake is to take them literally. (1992, p. 82)

This highlights the idea that there are distinct naturalizing projects at play here that should not be conflated with one another. But is this really true? In the section that follows, I will demonstrate that these projects cannot be teased apart from one another in this way. One cannot engage in one project without the other indirectly structuring, altering, and infecting its practices in an ineliminable and distortive way. This, in turn, greatly complicates both the methodological and ontological projects of naturalism.

3. Methodological and Ontological Naturalism intertwined

3.1 The infiltration of Methodological Naturalism into Ontological Naturalism

According to ON, being a useful posit in successful cognitive and neuroscientific theories is insufficient by itself to demarcate the robust cases of representation, from the instrumentally useful ascriptions of representation. As such, MN's criteria for realism are insufficient for the sort of robust realism they seek. And while this view is intuitive, it ultimately runs into the serious problem of assuming that there is a clear way of demarcating the robust cases of neural representations, from the instrumentally useful ascriptions of neural representations. But how do we do this?

Recall Searle's insistence that "there is nothing harmful, misleading, or philosophically mistaken about as-if metaphorical ascriptions [of intentionality]. The only mistake is to take them literally" (1992, p. 82). But given the pervasive and instrumentally valuable role that intentional/representational concepts play in successful neuroscientific, cognitive, and psychological theories, as well as everyday reasoning, it may be nigh impossible for us to tell apart which cases are supposed to be which. In this sense, keeping straight the literal from the metaphorical becomes exceedingly difficult since we may simply have no way to tell which cases in neuroscience and cognitive science meet the ontological naturalist's standards for robust representation and which do not.

To highlight the extent of this worry, consider that many scientists and philosophers disagree regarding which cases of intentional attribution in science are supposed to be the literal ones, and which are supposed to be the methodologically useful ascriptions. Carrie Figdor (2014, 2018), for instance, argues that if we examine the linguistic data regarding how intentional attributions are made in science, the evidence shows that scientists intend the attribution of representations literally when describing things like individual neurons, viruses, genes, and plants (among other things). Others have likewise noted the literal use of such concepts by scientists to describe things like immune systems (Colaço, 2025), and cyanobacteria (Bechtel & Bich, 2021).

Yet, many philosophers who engage in the ontological project of naturalizing representation are quick to deny that such descriptions ought to be treated literally (see, for example: Sarkar, 2000; Searle, 1980, 1992; Weber, 2005; Huebner, 2011; Mendelovici & Bourget, 2020). They insist that these are merely instances where representational or intentional descriptions are heuristically useful for scientists to invoke, but do not in fact identify robust representation or intentionality. The fact that scientists interpret such language literally just shows that they are being careless with their anthropomorphisms

(Searle, 1992, Weber, 2005, and Huebner, 2011 all make this claim). This is of course possible, but such a criticism would cut both ways, and calls into question whether we should be so confident when ascribing literal intentionality or representation to the human brain. To put it bluntly, how do we know we are not anthropomorphizing *ourselves*? As Walter Veit points out, "the seemingly 'undoubtable fact' that we have intentional content [may] itself [be] an illusion foisted upon us by the apparent intentional content of language" (2022, p. 38).

Without appealing to methodological naturalism and the explanatory success of scientific practice as our guide, how are we to demarcate which neuroscientific cases should be treated as the literal/robust cases of representation, and which should not? For the ontological project to get off the ground, we must already *know* which cases are which. To see why, consider that Searle (1980) famously insists that robust representation and intentionality only exists within biological systems. If we accept this view, then it immediately shapes the story we need to tell regarding how representations are instantiated by physical systems, and how they can be produced (since such a story must account for the representational differences between biological and non-biological systems). Conversely, if we believe that non-biological systems can indeed have robust intentionality and representation (e.g. Parisien & Thagard, 2008; Poldrack, 2021), then the story we need to tell will be radically different. Without any principled means of determining which cases of representation are the robust ones from the start, we will be building a metaphysical story of how representations are instantiated, and how they come about, so as to conform to the examples we *wish* to be true, instead of tracking the robust metaphysical phenomenon itself. Without methodological naturalism, ontological naturalists have no means of determining which cases of representation should count as the literal ones and why.

But perhaps there is another avenue that advocates of ON can pursue. As Rosa Cao puts it:

> To begin with, we are all intimately familiar with mental representations from the *inside*. Our own *thoughts* are representations, if anything is, and part of what makes the brain so fascinating as a target of inquiry is our rich suspicion that the familiar phenomenon and phenomenology of *thinking* is manifested in the brain in ways amenable to systematic scientific investigation. (2022, p. 151, emphasis in original)

Using this as our starting point, some have argued that we can appeal to our own introspection as a starting point to gain metaphysical insight into the robust cases of intentionality and representation. By directly observing our own mental representations, it can provide insights into representation in the brain without having to figure out from the start which cases of representation in scientific theorizing are those that identify the robust cases, as opposed to merely instrumentally useful ascriptions.

This provides a foundation upon which the ontological project of naturalization can be built that is not dependent on the methodological project. Angela Mendelovici (2018) explicitly goes this route, arguing that we can first identify representation and intentionality ostensibly by way of introspection, without needing to concern ourselves with how scientific (or even folk) theories posit intentional or representational phenomena. Her suggestion for understanding and identifying intentionality is…

> …to look past our descriptions of this phenomenon in terms of aboutness and related notions and focus instead on the phenomenon thus described. This is possible because we have a special access to this mental feature independent of any fuzzy or metaphorical descriptions: We can directly notice it through introspection, at least in some cases. (2018, p. 5)

She ultimately concludes from introspective observation that:

> When we introspectively notice intentional states, we notice the general phenomenon that we are tempted to describe as "directedness" or "saying something." But we also notice something we are tempted to describe as what our mental states are "directed at" or what they "say"; this is their (intentional) content. (2018, p. 8)

A similar idea has been defended by others as well (e.g. Crane, 2000; Pitt, 2011). If this is right, it allows us to identify the essential characteristics of robust intentionality and representation without worrying about which successful scientific theories to take as veridical, and which to not.

The problem with this seemingly intuitive story is that it assumes both that we can directly observe robust instances of intentionality or representation through introspection, and that we would know it when we saw it. In essence, it assumes that introspection is a kind of passive observation of the neurological and/or cognitive phenomena occurring in our head as they genuinely are. Yet there are good reasons to reject this.

Even if there are instances where someone notices the aboutness or directedness of their mental states through introspection, there is strong evidence that such noticings are directly structured and influenced by the scientific/folk theories that the subject already tacitly accepts, and linguistic categories they invoke, even if the subject doesn't consciously or explicitly bring them to bear on their introspective experience.

To illustrate how this might happen, let's consider someone who introspects on their experience of seeing a tree. Now imagine that when we ask this person if their introspection identifies an "of-ness" or "aboutness" to their experience, the person is at first utterly confused about what it is we are asking. They might, for instance, insist that they don't notice "aboutness"; what they notice is a tree. At which point we might explain to them that what they are noticing is in fact a conscious experience, and that this conscious experience must exist "inside their head" (since they might be hallucinating or dreaming that they see the tree). And so if it is this conscious experience that they are noticing, and they are noticing it “as a tree”, then clearly the experience must be an experience *of* a tree. This means there must be some internal state in their head that represents something else (namely a tree). And while the subject might concede to this line of reasoning after some coaxing, we are now already several degrees of abstraction away from the subject’s introspective report, let alone the experience itself, and have introduced all kinds of theoretical and linguistic concepts to help guide them to this conclusion. Armed with this new set of concepts and implicit theories, their introspection now conforms to such categories and they are now able to “notice” the aboutness when they introspect.

Put another way, it may seem as though we have a special kind of direct access to our internal mental states and the aboutness they possess, but this may be because we already talk in terms of things like “mental states”, “representations”, “content”, “experience”, and “aboutness”. In other words, the background theories and linguistic practices surrounding “representation” and “intentionality” structures and alters what we introspectively observe. As such, we cannot “look past our descriptions of this phenomenon in terms of aboutness and related notions and focus instead on the phenomenon thus described”, as Mendelovici claims, since the way we employ such descriptions implicitly frames and modifies how we perceive and understand ourselves and our own mental phenomena, whether we realize it or not.

There is indeed a great deal of evidence to support this. The background theories we adopt, language we use, and expectations we have, can influence our introspective observations in numerous different ways. Consider our introspective observation of our own emotions. Numerous scientific studies have shown that the emotion concepts we employ directly structure our understanding, introspection, and experience of our emotional states (Gendron et al., 2012; Lindquist et al., 2015; Barrett, 2006, 2017; Doyle & Lindquist, 2018; Hoemann et al., 2019; Fugate et al., 2020). Lisa Barrett, for instance, argues that many cross-cultural studies on emotion have shown that…

> …the way people learn about emotion categories and use conceptual knowledge determines what they see and feel. Variation in conceptualizing an instance of emotion, whether because of language use, context, culture, or individual differences in prior experience, will produce variation in which emotion is experienced and how it is experienced. (2006, p. 38)

And so the emotion concepts we already have in our repertoire, and theories we implicitly accept, will automatically change what our introspection tells us, and what our experiences of our emotions are like.

Or consider psychiatric disorders. The mere act of classifying someone as having a psychiatric disorder can change the way in which that person introspects, and alters what their subjective experiences are. This can result in them developing new symptoms they would not have developed had they not self-identified with the linguistic diagnostic category (see: Hacking, 1995, 1998; Kirmayer, 2005; Haslam, 2016).

Thus, the assumption that we merely passively observe our own mental phenomena when we introspect is not something we should be quick to accept. In fact, evidence suggests that the act of introspection itself is not a “looking inward” that gives us “special access” to our mental states (see: Churchland, 1996, p. 20; Schwitzgebel, 2008; Bayne & Spener, 2010; Spener, 2011; Carruthers, 2011; Cassam, 2014; Nanay, 2022). It may instead involve interpreting what we take our mental phenomena to be based on all kinds of physiological, behavioural, linguistic, environmental, and historical cues that we receive.

As a result, we cannot assume that we start with pure unfiltered observations of robust mental representation in the brain by way of introspection. Since the very act of observing our internal states *as* representations may itself already be heavily structured and influenced by the folk and scientific theories we tacitly accept. This means whether we accept, reject, or are ignorant of different neuroscientific, psychological, or cognitive theories that invoke representations will directly influence what we observe when we introspect, and thus will provide different answers as to which cases of representation we introspectively observe as genuine, and which we do not. In this respect, we cannot stop the methodological project of naturalizing representation from seeping into and distorting the ontological project of naturalizing representation, nor can we cleanly pull them apart.

3.2 The infiltration of Ontological Naturalism into Methodological Naturalism

Just as the methodological project infects and alters the ontological project, so too is it the case that the ontological project necessarily bleeds into, and distorts, the methodological project. Recall that according to MN, entities are natural and worthy of inclusion into our ontology if they are invoked within successful scientific theories and models. The question then becomes, how do we measure success? One criterion that is commonly considered essential for evaluating the success of our explanatory models and theories within cognitive science is in terms intervention. Scientific models and theories are explanatorily successful when they allow us to intervene on, manipulate, and control, phenomena in the world.

Consider, for instance, Chris Eliasmith who argues that successful explanations in the brain sciences are those which "provide a basis for both intervention in behaviour and the artificial reproduction of those behaviours" (2010, p. 316). Or Carl Craver, who argues that "explanatory models are much more useful than merely phenomenal models for the purposes of control and manipulation" (2006, p. 358). It is with this notion of scientific success in mind that philosophers and cognitive scientists have argued that models and theories which invoke mental representations have proven successful to science. Kriegeskorte & Diedrichsen (2019), for instance, tell us that "beyond the mere presence of the information in the neural activity, a representational interpretation implies that the information is used by downstream neurons in a way that contributes to behavior. We can test this hypothesis experimentally *by manipulating the activity and studying the effects on behavior*" (p. 408, my emphasis). Likewise, Nanay argues that "if we can manipulate mental representations in a way that would have direct influence on behavior, we would have a strong case for entity realism about mental representations" (2022, p. 82).

It is worth noting that it is the methodology of scientific practice (and not metaphysical arguments about how representational content is robustly constituted) that is supposed to justify the natural, and thus real, status of representations. Nanay, for instance, explicitly points out that his argument for realism regarding mental representation is "based on very specific empirical findings that we could only explain if we posit motor representations or pragmatic mental imagery." (2022, p. 89)

The problem with this sort of claim is that it is not as straightforward as it seems. This is because it is the ontological naturalist project that is used to frame and gauge these methodological successes. To illustrate, consider the study of place cells in the hippocampus. Place cells are standardly thought to play a role in the creation of cognitive maps; map-like representations that the brain employs for spatial navigation and orientation. Cognitive maps were originally posited by Edward Tolman (1948) to account for how rats were able to navigate mazes, and his findings are thought to have played a pivotal role in the birth of the cognitive revolution.

Since then, others have built upon his research to provide a more detailed account of how such cognitive maps are implemented in the brain, and how they are used by organisms to represent their environment (O'Keefe & Jonathan, 1971; O'Keefe & Conway, 1978; O'Keefe & Nadel, 1978; Hafner, 2000; Sullivan, 2010; Yuan et al., 2015; Bechtel, 2016). While this is the story that is traditionally told in cognitive science, the history is more complicated.

Evidence suggests that Tolman did not consider cognitive maps to be literal map-like representations encoded in the brain. As Tyler Delmore argues:

> Tolman's "cognitive map" was an attempt to provide an operational model of the mind that would allow psychologists to depict the extra-physiological variables unique to psychology's purview. On this re-interpretation, the map was not intended as a "map-like representation," held in the mind of an individual subject. It was a metaphor (and sometimes an actual diagram) for the causal relationships that psychology, and only psychology, could depict. What's important to emphasize is that, understood thusly, Tolman's promotion of maps did not require a turn to cognitivism. There were, I will show, already numerous instances of psychologists (including behaviorists) insisting on metaphorical "topologies," "spaces," "departments," and "economies" of behaviour. Tolman's maps, far from anticipating psychology's future, are best understood with respect to its past. (2024, p. 449)

Tolman's appeal to "cognitive maps" was used as a metaphor to highlight that internal variables within the organism were needed to account for behaviour beyond merely stimulus-response generalizations as some branches of behaviourism held. It was only much later, when cognitivism was in full swing, that the more common view of cognitive maps as literal map-like internal representations was developed (most famously by O'Keefe and colleagues. For example: O'Keefe & Jonathan, 1971; O'Keefe & Conway, 1978; O'Keefe & Nadel, 1978).

For Tolman, who worked broadly within the behaviourist tradition, certain entities were off the table metaphysically for him from the start when constructing his theory: internal mental representations (Delmore, 2024). And this structured how cognitive maps were invoked as posits within his theory, as well as the causal role they supposedly played (or didn't play) in the animal's behaviour. For O'Keefe and colleagues, however, cognitivism was in full swing, and representations (understood in terms of information about the environment encoded by the place cells themselves) were respectable entities to be invoked in theories in a way they were not for Tolman. This in turn shaped theory construction, experimental design, and the evaluation of successful interventions, in different ways. In this respect, what was methodologically naturalized given the success of Tolman and O'Keefe's theories were different given their different views regarding which ontologically naturalized entities they thought were acceptable to invoke in their theories from the outset.

To emphasize this point, let us consider a more contemporary version of this debate where the theorists involved are not separated by decades, or by the availability of distinct scientific tools. Consider current theories which describe the response properties of place cells in terms of representing distal features of the animal's environment. Such theories have proven undeniably successful in terms of prediction, explanation, and intervention. And so philosophers have good grounds to consider such representations to be naturalized according to the methodological project. But what exactly is it that has been naturalized? The metaphysical commitments scientists have regarding what counts as a respectable entity, and how representations are understood in light of this, from the outset effects how they develop their theories. This directly influences their understanding of what they have taken themselves to have naturalized, as well as directly feeds back into how they engage in, and understand, their methodological practices.

For instance, Francis Egan argues that positing representations is essential to our neuroscientific theories in domains like cognitive and computational neuroscience, but that representations should be understood as a linguistic gloss scientists apply when describing particular mechanisms. In her words:

> The cognitive characterization is essentially a gloss on the more precise account of the mechanism provided by the computational theory. It forms a bridge between the abstract, mathematical characterization that constitutes the explanatory core of the theory and the intentionally characterized pre-theoretic explananda that define the theory's cognitive domain. Unless the processes and/or structures given a precise mathematical specification in the theory

> are construed, under interpretation, as representations of such distal properties as edges, or joint angles, the account will be unable to address the questions that motivated the search for a computational theory in the first place, such questions as how are we able to see the three-dimensional structure of the scene from two dimensional images?, or how are we able to move our hand to grasp an object in sight? (Egan, 2010, p. 257)

Under this account, in virtue of being invoked by our best theories, MN would still license the inclusion of representations into our ontology; however, they would exist as something more akin to abstracta as opposed to specific physical-causal states of the brain itself (or a metaphysical relation between representation and represented). As she puts it, "it would be a mistake, though, to conclude that the structures posited in computational vision theories must (even in the gloss) represent their normal distal cause, and to find in these accounts support for a causal or information-theoretic theory of content" (p. 257). If scientists adopt this background metaphysical commitment, it directly influences what it is they think they are observing, predicting, and manipulating when it comes to representational theories of place cells.

With this interpretation in mind, the successful theories scientists invoke may allow them to manipulate the response properties of place cells, and predict certain behavioural outputs, but they would not be manipulating robust representational content *carried by* the place cell, since there is no such content to manipulate. By analogy, instead of modeling the response properties of place cells in terms of representational content, suppose scientists were to model them dynamically, as a vector moving through a state space. When doing so, it would be a mistake to think that the vector itself, as a mathematical object, is *carried by* the place cell in any sort of robust sense, or causally determining its behaviour. However, MN can still justify being realists about vectors as abstract objects, and scientists can manipulate the vector in their model by manipulating the response properties of the place cell. By adopting Egan’s background metaphysical commitments, representations, like vectors, would exist as abstracta for the methodological naturalist.

However, if scientists adopt a different set of metaphysical commitments that take place cells to be carrying representational content about the environment in a far more robust sense, then it changes not only how scientists understand the response rate of place cells, but also what sorts of interventions and experimental procedures would be considered important to carry out. For example, William Bechtel argues that “much neuroscience research is in fact directed at determining which neural processes are content bearers and understanding how they represent what they do. Content characterizations are not glosses on the research; the goal of the research is to determine what content the representations have.” (2016, p. 1291).

In support of this, Bechtel highlights how researchers attempt not just to understand the proximal mechanisms of the place cell (which Egan suggests representation-talk is used as a gloss to characterize), but instead what representational content the place cell in fact carries about its environment. By varying environmental factors, scientists attempted to understand what was, and was not, being represented by the response properties of the place cell. Moreover, they also attempted to discover how the representational content was being acquired by the place cell:

> The study of remapping has provided one of the main avenues for studying place cell representations in the hippocampus. The systematic changes in both the place fields and firing rates of place cells in response to changes in stimuli were pursued by the investigators in their attempt to determine how different properties of the vehicles, cell identity and its firing rate, encoded different information. The motivation for performing these and many other studies was to pin down exactly what changes in stimuli result in specific forms of place cell remapping. It is hard to understand efforts put into such research endeavors except on the assumption that

> researchers thought it important to determine how location in allocentric space is encoded in the activity of place cells and to identify the sources from which places cells acquire information about location. (2016, p. 1302)

Note how if scientists interpret representational content as robust, instead of a gloss on the neural mechanisms, then the sort of entity being methodologically naturalized by the same set of successful interventions are very different. More importantly, this in turn will have ramifications for how they engage in other experimental tasks with those metaphysical interpretations in mind. For instance, on this robust interpretation of representations, experiments to determine exactly what the content of the representation is, and how it is acquired by the system, become an important part of the research project. However, on Egan's account, there is no such content to find, and so no such experiments to conduct. As she puts, it, "an implication of the foregoing account of the role of representational content in computational models is that cognitive science has no need for a Naturalistic Semantics—the specification of nonintentional and non-semantic sufficient conditions for a mental state's having the meaning it does" (p. 257).

Or consider a radically different set of metaphysical commitments. One might grant that neuroscientific theories which invoke representations in characterizing the response properties of place cells are indeed successful, and thus are tracking something real, but deny that this "something" meets the criteria needed for being a representation. Segundo-Ortin & Hutto, for example, make this argument. They suggest that what scientists are tracking instead is a component part of the physical action itself:

> Crucially—focusing again on the parade case of place cells—the possibility that rat brains are using the forward-orientated firing of place cells for route planning is not the only available

> interpretation of the empirical evidence. Following Gallagher (2017), we contend that the fact that place cells fire in advance of action can be alternatively understood as "a constitutive part of the action itself, understood in diachronic, dynamical terms, rather than something decoupled from it" (Gallagher 2017, p. 14). On this view, anticipatory neural activity, operating on elementary timescales, can play a part in engendering larger-scale temporally extended cognitive activity. (2021, p. S20)

Whether scientists characterize the output behaviour of place cells in terms of representational content, or instead as part of the action of moving itself, has huge ramifications not only for what it is they take themselves to be manipulating in their successful experimental practices and theories of place cells, but also on how they evaluate their success, and what sorts of future research and methods count as legitimate and worthwhile. In this respect, the ON project directly shapes and modifies the MN project.

Now, the methodological naturalist might object here that theory interpretation is always present in theory construction, and this involves some metaphysical interpretation of the theory in question. As such, it is unsurprising that the methodological project presumes, and depends upon, discerning the commitments of the theory. But this in and of itself does not provide evidence that theorists must invoke metaphysical commitments independent of, or prior to, figuring out the role of a given representational posit in their theory.[3]

But such an objection is unsupportable. Scientists *must* begin with the ON project to frame their understanding of the phenomenon itself, and thus *are* required to invoke metaphysical commitments independent of, and prior to, figuring out the role of a given posit in their theory. These initial metaphysical commitments might change as their theories develop and they conduct new experiments,

[3] Special thanks to a blind referee for pushing me on this point.

but they must be present from the start and guide both research and the interpretation of experimental results. This is true of virtually all constructs in neuroscience and cognitive science.

For instance, Nedah Nemati highlights how this is unavoidable in neuroscientific practice, using studies of sleep in non-human animals as an example. She argues that neuroscientists must always begin with metaphysical assumptions regarding the phenomenon of sleep that they get from their own lived experience (which she called "Experientially Derived Notions", or EDNs), and that this is used as the foundation upon which theories of sleep, and methods of studying it, are built. She says:

> Here, researchers often have no choice but to start with what they know from firsthand experience. We not only have experiences of sleep (Thompson, 2015; Windt et al., 2016) but can also discern when someone or something else is asleep. To be a human among other humans and nonhuman animals, we must know when another person or animal is asleep and be able to distinguish this behavior from other states, such as unconsciousness or death. Thus, EDNs play a role in characterizing sleep as a behavior early on in research because, importantly, one cannot begin to test a behavioral construct without turning to some familiar notion of that behavior. In other words, to study sleep in animals is implicitly to study what happens to animals when they do *what I did last night when I fell asleep*. (2024, p. 4, emphasis in original)

Freek Oude Maatman (2021) similarly highlights this point when arguing that:

> Though often distal to our research questions, our base ontological assumptions about the nature of human cognition can also deeply influence our experiments. A most basic, common example is that we take the brain to be 'the seat' of the mind in some way, shape or form – but that without this commitment, neuroscience as a whole would cease to make sense, as well as reference to neuroscientific findings. Furthermore, allegiance to a paradigmatic position such as

> (radical) embodied cognition, radical behaviorism or computationalism can directly constrain the type of entities and mental processes we could use in psychological theory to begin with, by for example prohibiting talk of 'representations' (embodied cognition; see e.g. Shapiro & Spaulding, 2021), or requiring all processes to be specifiable as input-output algorithms (computationalism; see e.g. Piccinini, 2009) or as forms of behavior (radical behaviorism; see e.g., Skinner, 1953). Similarly, the complex systems approach's assumptions that the causal structure of cognition is interaction-dominant instead of component-dominant (Wallott & Kelty-Stephen, 2018; Van Geert, 2019) directly problematizes most existing psychological theories by effectively excluding the possibility of simple causal structures and isolatable entities or mechanisms. Such fundamental ontological commitments in turn also constrain how we conceive of psychological phenomena such as thought; e.g., as capacities that can be described as input-output algorithms (computationalist; Van Rooij & Baggio, 2020; 2021) or as processes instantiated in behavior that is inherently intertwined with its context (complex systems approach; Van Geert, 2019). (pp. 12-13)

Helen Longino likewise makes this argument regarding the study of phenomena like aggression. She notes that different scientific domains (behavioural genetics, neurobiology, developmental systems theory, etc) must adopt different metaphysical commitments regarding what aggression is from the start in order to set up experimental protocols that would allow each domain to study it using the tools and methods available to them. In doing so, these commitments determine what sorts of interventions and manipulations would be relevant to studying aggression *understood in one way as opposed to another*. And so while each can evaluate the success of their own interventions and manipulations, the different metaphysical interpretations of aggression from the start provide incompatible accounts of what the relevant interventions, manipulations, and causal variables are for studying it. As she puts it:

> Each approach employs methodologies that require particular ways of understanding the causal space. Some phenomena regarded as causally active in one approach are simply not included in another. These differential selections result in incongruous causal spaces. (2006, p. 118)

Different metaphysical commitments regarding what aggression is necessitate different sorts of interventions and manipulations needed to study it. And while we can evaluate whether the different set of interventions are methodologically successful, we cannot justify whether one *set* of interventions is better than another for studying aggression, unless we already have deep metaphysical commitments about what aggression as a phenomenon in the world is supposed to be (and thus which sorts of interventions will tell us whether we are tracking it). In this regard, scientists *must* adopt robust metaphysical commitments regarding the phenomena being studied from the outset, and these commitments are not unconstrained. They are directly structured by the sorts of entities already deemed respectable by the working scientist.

While one metaphysical interpretation of representation can justify setting up interventions that can confirm/disconfirm theories *with that metaphysical view of representation in mind*, such criteria would be unhelpful for confirming/disconfirming theories *with a different metaphysical view of representation in mind*. In this way, MN cannot answer questions about whether one should understand the place cell in terms of representations *qua abstract object, qua robust phenomenon in the brain,* or *qua instrumental characterizations of some component of a complex action*. Each metaphysical interpretation brings with it distinct sets of criteria for determining the success of the interventions and manipulations needed to support them, with no shared set of criteria for what should count as successful interventions *across* interpretations in order to support one over the others. As Helen Longino puts it:

> Each approach is characterized by its own criteria and standards for evaluating particular studies carried out within its framework. It is not clear that there is or could be a common empirical standard by which to evaluate each approach as against the others, and participants in the debates point to different kinds of features of their approaches as evidence of superiority to those of their rivals. (2001, p. 691)

Yet scientists *must* take a metaphysical stand on such issues from the outset, since otherwise they have no way to constrain theory construction, or determine experimental protocols for studying phenomena, let alone evaluating their successes. Their pre-existing views of what acceptable naturalistic entities are, and what the phenomenon in nature they wish to study is metaphysically taken to be, must begin prior to understanding the variables in the theories they construct which makes reference to them (see: Hochstein 2019).

This idea is further supported by Jacqueline Sullivan's (2010) account of how different metaphysical commitments regarding representations can, and do, influence neurobiological theory construction and experimental practice:

> In light of the aforementioned considerations, I want to draw a distinction between representations playing a minimal role versus a substantive role in the contexts of experimentation and explanation in cognitive neurobiology. [...] In cases in which representation plays a minimal role in the context of experimentation, an investigator operationally defines a form of learning in terms of observable changes in behavior, and those changes constitute the targets of explanation. He then intervenes in the activity of cells and molecules, determines the effects on behavior, and explains those effects in terms of the cellular and molecular

> interventions. […] In contrast, in cases in which representations play a substantive role in cognitive neurobiology, changes in internal representations may be construed as the phenomena that one aims to produce and then intervene in or disrupt by undertaking a molecular intervention. In this case, representations would likely be the explanatory targets of a mechanistic explanation. (p. 881)

What this means is that the methodological project of naturalism is necessarily and irrevocably intertwined with the ontological project of naturalism since we evaluate and gauge our methodological successes in light of our ontological and metaphysics commitments. The two projects must go hand in hand, since to engage in one requires that we simultaneously engage in the other (whether we intend to or not). And so we are left with two projects that cannot be cleanly distinguished or pulled apart from each other, despite seemingly having distinct goals and motivations.

4. Can the two projects mutually support each other?

One might think that the consequences of all this are not particularly worrisome. Why not let the different projects inform one another? Since the different projects have different aims and goals, we can use each project as a way of providing independent checks and balances on the other. The ontological project can provide support for, or arguments against, the findings of the methodological project, and vice versa. And so we have two different projects that inform each other to progress and improve. Morgan & Piccinini (2018) seem to have something like this in mind when they claim that:

> This strategy seemed to provide an appealing division of labor for solving the puzzles of intentionality from within the purview of natural science: Cognitive scientists would construct empirical theories of the mental representations that explain cognitive capacities, while

> philosophers would articulate a set of conditions, expressed in naturalistically respectable terms, that determine the semantic content of those representations. (p. 125)

The problem with this account is that each project of naturalism does not so much provide distinct means of constraining or supporting the other, so much as reflect back the assumptions that each project has already implicitly taken from the other.

Let's return to our example of cognitive maps. If we adopt Tolman's theory of cognitive maps as the successful empirical theory which is to act as the support or constraint on the ontological project, then the success of such a theory would justify the natural, and thus real, status of cognitive maps *as abstract objects*. They would not license the further claim that such maps were actual map-like representations encoded in the brain itself (as Tolman's theory denies such an interpretation). And so there would be no set of conditions, expressed in ontologically robust naturalistically respectable terms, that determine the semantic content of those cognitive maps. Conversely, if we choose O'Keefe's theory as the successful empirical theory, then there *would* be a set of conditions, expressed in ontologically robust naturalistically respectable terms, that determine the semantic content of those cognitive maps.

But notice, the problem is that each of these two theories already have baked into them deep metaphysical commitments of what cognitive maps are, and how they are/are not instantiated in the brain, based on the entities that they already deemed respectable from the start. And so such theories do not provide independent support or constraints on the ontological project, they just reflect back the assumptions that such theories already took from the ontological project when they were developed. The same set of interventions on the animal's behaviour, and on the response properties of the place cells, fit with both the interpretation that cognitive maps are ontologically robust representations encoded in the brain *and* that they are abstract objects denoted by scientific theories necessary for us to understand the place cell's response properties and the animal's behaviour. The two theories adopt different ontological commitments as to what representations, and cognitive maps, are supposed to be in order to frame and understand their interventions and practices, and they only provide support for, or constraints on, the ontological project in virtue of already baking such ontological commitments into their account.

Similarly, we cannot argue that one theory is better than the other because one better fits the naturalized metaphysics of the ontological project, since again the question of what the respectable metaphysical entities are is based on which scientific theory we have liberally borrowed from. If we use Tolman as a guide for what entities count as “respectable” or not for our ontological project, then we have grounds to reject O’Keefe’s theory for failing to conform with the respectable metaphysics (in virtue of positing map-like representations in the brain, which are not acceptable ontological entities). Conversely, if we use O’Keefe’s theory as our guide for what entities count as “respectable”, then robust map-like representations will be acceptable to include into our metaphysical story of how representations are constituted.

In order to articulate a set of conditions, expressed in naturalistically respectable terms, that determine the semantic content of the representations invoked by successful neuroscientific theories, it first requires determining which scientific theories to appeal to. And different successful neuroscientific theories bake different implicit metaphysical assumptions as to what representations are into their account in order to determine what the best methods for studying them are. And so we will be articulating a set of conditions, expressed in naturalistically respectable terms, that conform to the metaphysics already presupposed by the theory we happen to tacitly accept. And so it cannot provide genuine confirming or discomforting accounts to help support or undermine the ontological theory in question.

In the case of Tolman versus O’Keefe, one might be tempted to point to the fact that O’Keefe, in developing his account well after Tolman, had better means of providing interventions and manipulations to support his theory, giving the methodological edge to O'Keefe's theory (and suggesting that cognitive maps are indeed robust entities/processes in the brain). But as we've seen, all the interventions and predictions carried out by O'Keefe in support of his theory are also compatible with a Tolman-like-interpretation of cognitive maps as useful metaphors to characterize complex coordinated mechanistic activities (and not robust map-like representations encoded in the brain itself). In other words, O'Keefe's interventions and predictions are just as compatible with the idea that cognitive maps are merely heuristically useful glosses on mechanisms (a la Egan), or as heuristically useful ways of characterizing part of a complex behaviour in action (a la Segundo-Ortin & Hutto). In this regard, the methodological project cannot provide constraints on which of these accounts the ontological naturalist should adopt. Only if the different interpretations bring with them different sets of predictions *of the same behaviours,* or predict *different outcomes for the same set of manipulations*, can the methodological naturalist justify the natural status of the entities of one interpretation over the other. Yet all three interpretations fit with the same interventions and predictions. While each account may propose *novel* sets of manipulations and interventions for accounting for their own unique interpretations, those interventions would only be informative given that interpretation, not across interpretations.

Put simply, the two projects of naturalism interact at such a basic level that we can't engage in the methodological project of theory construction, application, and verification, without first engaging in the ontological project of determining what sorts of entities we should consider metaphysically acceptable to invoke, how we metaphysically understand the phenomenon in the world prior to theory construction, and what we think their constitution entails. Conversely, we can't determine whether an entity is constituted by respectable ontological elements without simultaneously engaging in the methodological project of using our best theories as a guide for what sorts of elements are acceptable to use or not. The two projects are so deeply intertwined that we can't use one project as a means of justifying the other since the two projects already help themself to features of the other in their most basic functioning. This means a problematic ontological assumption can undermine a methodological justification for an entity, and a problematic methodological assumption can undermine the ontological justification for an entity.

5. Conclusion

In this paper I have argued that the common assumption that there are distinct projects of naturalization when it comes to representation in neuroscience and cognitive science is deeply problematic, and has further complicated the debates surrounding neural representation. MN and ON must go hand-in-hand at every step, and so the assumption that we can bracket one project to focus on the other must be met with suspicion. Conversely, the assumption that we can use each project as a means of providing independent confirmation or disconfirmation for the other must likewise be met with equal suspicion. Current disputes regarding the naturalization of representation in neuroscience and cognitive science must be viewed through this lens.

It is worth stressing that my intention is not to suggest that this in any way undermines naturalism, or the naturalization of representation in neuroscience and cognitive science. My intention instead has been to highlight that the criteria for determining if representations are, or are not, natural (and thus real) are not nearly as straightforward or clearcut as current debates assume.

For instance, Segundo-Ortin & Hutto (2021) are right to criticize Thomson & Piccinini (2018)'s claim that scientists directly observe and manipulate representations in the laboratory by pointing out that they can accept that they are observing and manipulating something, but that this something may not be a representation at all, and only seems so if they bake various metaphysical commitments into their view from the outset. However, Morgan & Piccinini (2018) are equally right to criticize Segundo-Ortin & Hutto for making the same type of mistake:

> Second, and more problematically, such anti-representationalist arguments tend to target a stereotype of representations as static, word-like symbols, without clearly identifying fully general conditions for something to count as a representation. Representations might very well be action-oriented, dynamical, or both. (p. 129)

Just like Thomson & Piccinini, Segundo-Ortin & Hutto assume a particular metaphysical view of what representational content must be in order to argue that the scientific successes of theories which invoke representations do not validate them. And yet there are many metaphysical views of representation endorsed by both philosophers and cognitive scientists that do not conform to their criteria. And so their insistence that a non-representational theory of the mind best accounts for the metaphysics runs into the same problem.

This means that genuine progress in these debates will not occur until both sides take seriously the complex and subtle ways in which metaphysics and methodology are baked into each other at every step of inquiry. A great deal more care must be taken within philosophy of neuroscience and cognitive science regarding the interwoven nature of methodological and ontological naturalism before disputes about the natural status of representations can move forward.

References

Barrett, L. (2006). Solving the Emotion Paradox: Categorization and the Experience of Emotion. *Personality and Social Psychology Review*, *10*(1), 20-46. doi: 10.1207/s15327957pspr1001_2

Barrett, L. (2017). *How emotions are made*. Houghton Mifflin Harcourt.

Bayne, T., & Spener, M. (2010). Introspective humility. *Philosophical Issues, 20*, 1–22. https://www.jstor.org/stable/41413543

Bechtel, W. (2016). Investigating Neural Representations: The Tale of Place Cells. *Synthese*, *193*(5), 1287–1321. https://www.jstor.org/stable/43921177

Bechtel, W., & Bich, L. (2021). Grounding cognition: heterarchical control mechanisms in biology. *Phil. Trans. R. Soc. B*, (376), 20190751. https://doi.org/10.1098/rstb.2019.0751

Cao, R. (2022). Putting representations to use. *Synthese*, *200*(2), 151. https://doi.org/10.1007/s11229-022-03522-3

Cao, R. (2012). A teleosemantic approach to information in the brain. *Biol Philos*, *27*, 49–71.

Colaço, D. (2025). On Consistently Assessing Alleged Mnemonic Systems (or, why isn't Immune Memory "really" Memory?). *Review of Philosophy and Psychology*, 1-19. https://doi.org/10.1007/s13164-025-00768-x

Cassam, Q. (2014). *Self-knowledge for Humans*. Oxford University Press.

Churchland, P. (1996). *The Engine of Reason, the Seat of the Soul*. Bradford and MIT Press.

Collins, J. (2015). Naturalism without metaphysics. In E. Fischer & J. Collins (Eds.), *Experimental Philosophy, Rationalism, and Naturalism: Rethinking Philosophical Method* (pp. 85-109). Routledge.

Crane, T. (2000). Introspection, Intentionality, and the Transparency of Experience. *Philosophical Topics*, *28*(2), 49-67.

Craver, C. (2006). When mechanistic models explain. *Synthese*, *153*(3), 355–376.

Carruthers, P. (2011). *The opacity of mind: An integrative theory of self-knowledge*. Oxford University Press.

Delmore, T. (2024). Re-Charting Tolman's Cognitive Maps. *Dialogue: Canadian Philosophical Review/Revue canadienne de philosophie*, *63*(3), 447-466.

Dennett, D. (1991). Real Patterns. *The Journal of Philosophy*, *88*, 27-51.

Downey, A. (2018). Predictive processing and the representation wars: a victory for the eliminativist (via fictionalism). *Synthese*, *195*, 5115–5139.

Doyle, C. & Lindquist, K. (2018). When a word is worth a thousand pictures: Language shapes perceptual memory for emotion. *Journal of Experimental Psychology: General*, *147*(1), 62–73.

Dretske, F. (1981). *Knowledge and the Flow of Information*. MIT Press.

Dretske, F. (1988). *Explaining Behavior: Reasons in a World of Causes*. MIT Press.

Egan, F. (2010). Computational models: A modest role for content. *Studies in History and Philosophy of Science Part A*, *41*(3), 253-259.

Eliasmith C. (2010). *How we ought to describe computation in the brain*. Stud Hist Philos Sci Part A, *41*, 313–320

Eliasmith, C. (2012). The Complex Systems Approach: Rhetoric or Revolution. *Topics in Cognitive Science*, *4*, 72-77.

Favela, L. (2021). The Dynamical Renaissance in Neuroscience. *Synthese*, *199*, 2103-2127.

Favela, L. & Machery, E. (2023). Investigating the concept of representation in the neural and psychological sciences. *Frontiers in Psychology*, *14*, 1165622. doi: 10.3389/fpsyg.2023.1165622

Figdor, C. (2014). On the Proposed Domain of Psychological Predicates. *Synthese* 194, *11*, 4289-4310.

Figdor, C. (2018). *Pieces of Mind: The Proper Domain of Psychological Predicates*. Oxford University Press.

Flanagan, O. (2006). Varieties of naturalism. In P. Clayton & Z. Simpson (Eds.), *The Oxford Handbook of Religion and Science* (pp. 430-452). Oxford University Press.

Fodor, J. (1987). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. MIT Press.

Fugate, J., MacDonald, C., & O'Hare, A. (2020). Emotion words' effect on visual awareness and attention of emotional faces. *Frontiers in Psychology*, *10*, 2896.

Gallagher, S. (2017). *Enactivist interventions. Rethinking the mind*. Oxford University Press.

Gendron, M., Lindquist, K., Barsalou, L., & Barrett, L. (2012). Emotion words shape emotion percepts. *Emotion*, *12*(2), 314–325.

Gładziejewski, P. & Miłkowski, M. (2017). Structural representations: Causally relevant and different from detectors. *Biology & Philosophy*, 32, 337-355.

Hacking, I. (1995). The looping effect of human kinds. In D. Sperber, D. Premack, & A.J. Premack (Eds.), *Causal Cognition: A Multidisciplinary Debate* (pp. 351–394). Clarendon.

Hacking, I. (1998). *Mad Travelers: Reflections on the Reality of Transient Mental Disease*. University Press of Virginia.

Hafner, V. (2000). Cognitive maps for navigation in open environments. *Proceedings 6th international conference on intelligent autonomous systems (IAS-6)* (pp. 801–808). IOS Press.

Haslam, N. (2016). Looping effects and the expanding concept of mental disorder. *Journal of Psychopathology*, *22*, 4-9.

Hochstein, E. (2019). How Metaphysical Commitments Shape the Study of Psychological Mechanisms. *Theory & Psychology*, *9*(5), 579-600.

Hoemann, K., Xu, F., & Barrett, L. (2019). Emotion words, emotion concepts, and emotional development in children: A constructionist hypothesis. *Developmental Psychology*, *55*(9), 1830–1849.

Huebner, B. (2011). Minimal Minds. In T. L. Beauchamp and R. G. Frey (Eds.), *The Oxford Handbook of Animal Ethics* (pp.441-468). Oxford University Press.

Hutto, D. & Myin, E. (2013). *Radicalizing Enactivism: Basic Minds without Content*. MIT Press.

Hutto, D. & Myin. (2014). 'Neural Representations Not Needed – no More Pleas, Please,' *Phenomenology and the Cognitive Sciences, 13*(2), 241–256.

Kirmayer, L. (2005). Culture, Context and Experience in Psychiatric Diagnosis. *Psychopathology*, *38*(4), 192–196.

Lindquist, Kristen, Jennifer MacCormack, and Holly Shablack. 2015. "The role of language in emotion: Predictions from psychological constructionism." Frontiers in Psychology 6: 444.

Longino, H. E. (2001). What do we measure when we measure aggression?. *Studies in History and Philosophy of Science Part A*, *32*(4), 685-704.

Longino, H. (2006). Theoretical pluralism and the scientific study of behavior. In S. Kellert, H. Longino, & C. K. Waters (Eds.), *Scientific pluralism* (pp. 102–132). University of Minnesota Press.

Maddy, P. (2011). *Defending the Axioms*. Oxford University Press.

Martinez, J. E. (2025). Facecraft: Race reification in psychological research with faces. *Perspectives on Psychological Science*, *20*(1), 182-194.

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

Mendelovici, A. & Bourget, D. (2014). Naturalizing Intentionality: Tracking Theories Versus Phenomenal Intentionality Theories. *Philosophy Compass*, *9*(5), 325-337.

Mendelovici, A., & Bourget, D. (2020). Consciousness and Intentionality. In U. Kriegel (Ed.), *The Oxford Handbook of the Philosophy of Consciousness*. Oxford University Press.

Millikan, R. (1984). *Language, thought and other biological categories*. MIT Press.

Millikan, R. (1989). Biosemantics. *The Journal of Philosophy*, *86*(6), 281-297.

Morgan, A., & Piccinini, G. (2018). Towards a cognitive neuroscience of intentionality. *Minds and Machines*, *28*, 119-139.

Nemati, N. (2024). Rethinking Neuroscientific Methodology: Lived Experience in Behavioral Studies. *Biological Theory*, *19*(3), 184-197.

Nanay, B. (2022). Entity Realism About Mental Representations. *Erkenn*, *87*, 75–91.

Oude Maatman, F. (2021). Psychology's theory crisis, and why formal modelling cannot solve it. *PsyArXiv*. https://psyarxiv.com/puqvs/

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely moving rat. *Brain Research*, *34*, 171–175.

O'Keefe, J., & Conway, D. (1978). Hippocampal place units in the freely moving rat: Why they fire where they fire. *Experimental Brain Research*, *31*, 573–590.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford University Press.

Parisien, C., & Thagard, P. (2008). Robosemantics: How Stanley the Volkswagen Represents the World. *Minds and Machines*, *18*, 169–178.

Piccinini, G. (2009). Computationalism in the Philosophy of Mind. *Philosophy Compass*, *4*, 515-532

Piccinini, G. (2018). Computation and Representation in Cognitive Neuroscience. *Minds and Machines*, *28*, 1–6.

Pitt, D. (2011). Introspection, Phenomenality, and the Availability of Intentional Content. In T. Bayne & M. Montague (Eds.), *Cognitive Phenomenology* (pp. 141-173). Oxford University Press.

Poldrack, R. (2021). The Physics of Representation. *Synthese*, *199*, 1307-1325.

Richmond, A. (2024). What is a theory of neural representation for? *Synthese*, *205*(1), 14.

Rosenberg, A. (2015). The genealogy of content or the future of an illusion. *Philosophia*, *43*(3), 537–547.

Sarkar, S. (2000). Information in Genetics and Developmental Biology: Comments on Maynard Smith. *Philosophy of Science*, *67*(2), 208-213.

Schwitzgebel, E. (2008). The Unreliability of Naive Introspection. *Philosophical Review*, *117*, 245–273.

Searle, J. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, *3*, 417-457.

Searle, J. (1992). *The Rediscovery of the Mind*. The MIT Press.

Segundo-Ortin, M., & Hutto, D. (2021). Similarity-based cognition: radical enactivism meets cognitive neuroscience. *Synthese*, *198*(Suppl 1), 5–23.

Shapiro, L. & Spaulding, S. (2021). Embodied Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition). https://plato.stanford.edu/archives/win2021/entries/embodied-cognition/

Shapiro, L. (1997). The nature of nature: Rethinking naturalistic theories of intentionality. *Philosophical Psychology*, *10*(3), 309-322.

Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.

Skinner, B.F. (1953). *Science and human behavior*. Macmillan

Spener, M. (2011). Using first person data about consciousness. *Journal of Consciousness Studies*, *18*, 165–179.

Stalnaker, R. (1987). *Inquiry*. Bradford Books.

Stewart, J., Gapenne, O., & Di Paolo, E. (2010). *Enaction: Toward a New Paradigm for Cognitive Science*. MIT Press.

Sullivan, J. (2010). A role for representation in cognitive neurobiology. *Philosophy of Science*, *77*(5), 875-887.

Taylor, S. (2022). Cognitive Instrumentalism about Mental Representation. *Pacific Philosophical Quarterly*, *103*, 518–550.

Thomson, E., & Piccinini, G. (2018). Neural Representations Observed. *Minds and Machines*, *28*, 191–235.

van Geert, P. L . (2019). Dynamic systems, process and development. *Human Development*, *63*(3–4), 153–179.

van Rooij, I., & Baggio, G. (2020). Theory Development Requires an Epistemological Sea Change, *Psychological Inquiry*, *31*(4), 321-325.

Veit, Walter. (2022). Revisiting the Intentionality All-Stars. *Review of Analytic Philosophy*, *2*(1), 31-53.

Wallot, S. & Kelty-Stephen, D.G. (2018). Interaction-Dominant Causation in Mind and Brain, and Its Implication for Questions of Generalization and Replication. *Minds and Machines*, *28*, 353–374

Ward, Z. (2023). Muscles or Movements? Representation in the Nascent Brain Sciences. *Journal of the History of Biology*, *56*(1), 5-34.

Weber, M. (2005). Genes, Causation and Intentionality. *History and Philosophy of the Life Sciences*, *27*, 407–420.

Weiskopf, D. (2017). The Explanatory Autonomy of Cognitive Models. In D. M. Kaplan (Ed.) *Explanation and Integration in Mind and Brain Science* (pp. 44-69). Oxford University Press.

Williams, D. (2018). Predictive Processing and the Representation Wars. *Minds and Machines*, *28*, 141–172.

Yuan, M., Tian, B., Shim, V.A., Tang, H., & Li, H. (2015). An entorhinal-hippocampal model for simultaneous cognitive map building. *Proceedings of the twenty-ninth AAAI conference on artificial intelligence*. *29*(1)