Fictionalism about Neural Representations

1. Introduction

This paper articulates and explores a novel form of Mental Fictionalism: Fictionalism about the *neural representations posited by cognitive science*. Cognitive science appears to be committed to neural representations. These representations are claimed to be the springs of our thought and action: they drive our behaviour, determine our thoughts, memories, and inferences. However, despite the central role of neural representations in cognitive science, it is hard to explain what is meant by 'representation' in a way that does not incur problematic commitments. The representations in question clearly cannot be *conventional* representations that gain their representational content and status through our social conventions; for we are rarely aware that such representations exist, and no adequate social conventions regarding them appear to be in play. The standard reply is that neural representations are representations of a different sort: *original* or *natural* representations. This class of representations gain their representational status independently of, and in some sense prior to, our social conventions. But what is a *natural representation*? Attempts to answer this question—*naturalising representation*—have been on-going since the 1970s. Unfortunately, this project to date has been largely unsuccessful. Many contemporary theorists are sceptical that an adequate naturalistic theory of representation will ever emerge.

For this reason, some theorists have been drawn to *Eliminativism* about talk of neural representations in cognitive science.[1] If cognitive science could stop appealing to neural representations, then there would be no need to give an account of representation, and therefore no need to give a naturalistic account. However, thoroughgoing Eliminativism about neural-representation talk is a hard road to follow. Although some cognitive phenomena can be explained in nonrepresentational terms, other aspects of cognition appear stubbornly resistant to nonrepresentational

"Fictionalism about Neural Representations" by Mark Sprevak,
The Monist, vol. 96, no. 4, pp. 539–560.

explanation. Attributing neural representations seems to be the best way to explain many of our cognitive abilities.[2]

Cognitive science appears to face a dilemma: *either* it uses neural-representation talk and is lumbered with the task of naturalising representation, *or* a radical and undesirable revision to the practice of cognitive science is required.

Neural Representation Fictionalism (NRF) offers a neat way out. NRF opens up a third option: allowing us to *use* neural-representation talk in cognitive science *without* the cost of naturalising representation. NRF promises to rid us of one of the biggest problems facing representation talk in cognitive science without the pain required by Eliminativism. NRF purports to deliver the benefits of both Realism and Eliminativism with the costs of neither. The only downside of NRF is that it would require us to reinterpret neural-representation talk in cognitive science in a fictionalist way. At least on the face of it, it is not obvious that this is not a price worth paying. NRF seems worth exploring.

Fictionalism about a given discourse is the view that claims *C* in that discourse involve genuine statements of fact—they aim to describe the world—but, in contrast to Realism, those claims *C* do not aim at truth. Instead, they serve some other purpose. A fictionalist might remain *agnostic* about the truth value of her claims *C* (as van Fraassen [1980] does), or she may declare that the claims *C* are *literally false* in spite of their cognitive value (as do Nolan, Restall, and West [2005]).

Fictionalism of many stripes has become popular in recent years. Forms of Fictionalism have been developed for mathematical discourse, moral discourse, modal discourse, and negative existential talk. In each case, the motivation bears a striking resemblance to the problem facing cognitive science above. We have a practice—mathematical talk, moral talk, modal talk, or negative existential talk—that appears to commit us to the existence of troublesome entities—*numbers*, *moral facts*, *possible worlds*, *nonexistent objects*. Attempts to explain how these entities fit into the physical world face serious challenges. However, dispensing with the entities by eliminating talk of them is not attractive either, since talking about them appears essential for us to achieve our ends. According to the Fictionalist, the relevant discourse should be understood as a fiction; the fiction plays a similar cognitive role in our lives to that which was thought to be played by the truth, but it does not carry a commitment to the existence of the relevant entities. Fictionalism would allow us to talk like a Realist but without incurring the ontological commitments.

NRF is an instance of a broader form of Fictionalism: *Mental Fictionalism*. Psychological talk is rife with troublesome ontological commitments and seems to be potentially fertile ground for developing forms of Fictionalism. Demeter (2009a; 2009b; 2010) and Wallace (2007) explore Fictionalism about Folk Psychology. On their view, Folk Psychology is false, but useful for prediction, understanding, and normative evaluation. Dennett (1991; 2002) develops a form of Fictionalism *avant la lettre* about talk of conscious experience. According to Dennett, talk of conscious experience is false but useful because it allows us to talk concisely about complex disjunctions of physical properties such as colours. Dennett (1991) also defends a form of Fictionalism about the self: such talk serves a useful purpose by allowing us to better plan and coordinate our action.

My intention in this paper is to explore Fictionalism about neural-representation talk in cognitive science. This by necessity can only be a preliminary scouting, but it at least indicates the general shape the view must take and the main challenges that it faces. The plan for the paper is as follows. In Section 2, I introduce the general Fictionalist framework. In Section 3, I describe the kind of neural-representation talk in cognitive science that NRF will target. In Section 4, I state NRF and canvas NRF's benefits. In Section 5, I argue that, despite its benefits, NRF faces two serious objections. These objections are that: (1) NRF does not, in the end, avoid the problem of naturalising representation; (2) NRF cannot adequately serve cognitive science because fictional neural representations cannot play the explanatory role required by cognitive science.

2. Fictionalism

The central notion of Fictionalism is that of *ontological commitment*.[3] Ontological commitment is the idea that some of our practices, and, in particular, our linguistic practice when we give our best description of the world, commit us to the existence of certain entities. If our best description of the world contains reference to *X*s, then we are committed, on the face of it, to the existence of *X*s. Correspondingly, if we want to know which entities exist, we should start by looking at the ontological commitments of our best theories.

The notion of ontological commitment has been hugely influential in contemporary metaphysics. The precise details of the notion are controversial, but the basic idea—that our best description of the world commits us to an ontology—is widely accepted.

If one accepts this basic idea, problems quickly begin to emerge. Consider the following claims:

(1) $2 + 2 = 4$

(2) Torture is wrong.

(3) The course of biological evolution could have been different.

(4) Phlogiston does not exist.

We typically take all these claims to be *fact stating* and *true*. Claims (1–4) are uttered with all sincerity, and they figure in, or are entailments of, our best descriptions of the world. Yet if (1–4) are both fact stating and true, then on the face of it they commit us to a range of entities: *numbers*, *moral facts*, *modal facts*, and *nonexistent entities*. These entities pose a number of well-known and daunting challenges, including explaining how the entities fit into a physical world, how agents like ourselves can have knowledge of them, and how agents like ourselves can refer to them. These problems have prompted many philosophers to doubt whether a Realist interpretation of the relevant discourse is the best strategy. An alternative is Eliminativism: one may say that (1–4) should *not* be asserted by our best theories. Talk of numbers, moral facts, modal facts, and nonexistent entities should be eliminated from serious fact-stating talk. However, Eliminativism also faces well-known problems. It is hard to eliminate mathematical, moral, modal, or negative existential talk while still allowing us to achieve our ends.

Fictionalism appears to offer an easier way out. The Fictionalist starts from the observation that we often engage in serious fact-stating talk without incurring ontological commitments. Common examples come from figurative language. As Yablo (1998, 233) says, "not even Quine considers it ontologically committing to say in a *figurative* vein that there are *X*s." A figurative assertion may be fact stating (e.g., 'Nothing gets my goat as much as chewing gum in class') without the ontological commitment that a literal assertion would normally bring (i.e., no commitment to the existence of a *goat*). Figurative language can be understood to include a range of linguistic devices that use fact-stating language for non-truth-stating ends, including metaphor, hyperbole, pretence, and supposition.

A notable property of figurative language is that even if figurative claims are false, they may nonetheless possess significant cognitive value. One value of figurative language is that it provides an alternative way of expressing claims than literal assertion. But the benefits of figurative language go beyond merely being an alternative mechanism for expressing claims that could have been stated more plainly. Figurative language often functions as a powerful inferential device: it prompts us to engage in a range of appropriate and useful inferences (e.g., 'Jimi is on fire today' primes a range of inferences about how Jimi will perform and react). Figurative language can provide a concise description where a literal description is too long-winded or simply unavailable (e.g., 'she gave him a piercing glance'). Figurative language can provide us with models with which to make sense of the world (e.g., Romeo saying that "Juliet is the sun" provides us with a model to understand Romeo's thoughts and behaviour). Figurative language can simplify the world in helpful ways by the use of pretence (e.g., 'water is an incompressible fluid'). Figurative language is also a particularly apt fit for human psychology; it tends to 'stick' in our minds and inspire future enquiry in a way that literal language rarely achieves (e.g., 'the clockwork universe'). Figurative language allows us to describe complex situations concisely in a way that is pregnant with appropriate inferences and understanding. It is not difficult to see why this is something that we value. The cognitive virtues of figurative language may outweigh the vice of introducing literal falsehood into our best description of the world.

3. Neural Representations

Neural representations play an important role in cognitive science. Cognitive science often ascribes neural representations when certain brain regions or individual neurons of an animal are demonstrated to respond strongly, and selectively, to certain stimuli. Activity in those brain regions or individual neurons is taken to *represent* those stimuli. The representations in question are *subpersonal*: they are attributed to parts of the animal (brain regions and neurons), not to the animal as a whole (e.g., as beliefs, desires, and thoughts would be). The neural representations are often *nonconscious*: the animal is unable on the basis of conscious reflection alone to know that it has them. The neural representations are also often assumed to be more fundamental than, and to somehow *ground*, the animal's con-

scious and personal-level thought. We will return to this grounding claim in Section 5. What concerns us here is that—independent of its use in a grounding claim—neural-representation talk is used by cognitive science to explain animal behaviour.

Cognitive science appeals to the fact that an animal represents its environment using neural activity to explain why the animal succeeds in its environment. Neural representations that are appropriate to an animal's environment explain why the animal is successful. The animal is successful because it consults an internal model that predicts what will happen in the environment, and the model guides the animal's action accordingly. Neural representations and the animal's subpersonal inferences over those neural representations explain the animal's success. Just as the use of a map of the London Underground explains the success of a visitor to London in navigating around the city, so the use of appropriate neural representations explains the success of an animal in dealing with its environment. Unsuccessful behaviour can be explained by the animal having the wrong neural representations. If false or inappropriate neural representations are deployed by the animal, one would expect systematic mistakes in behaviour. Numerous errors in animal behaviour can be rendered comprehensible if one understands the animal acting 'as if' a stimulus were present. Similarly, systematic wrong turns taken by a visitor to London can be explained if we discover that the visitor is using an incorrect map.

The role of neural representations is not confined to explaining behaviour. Neural representations also play a role in explaining off-line cognitive phenomena such as memory, imagery, anticipation, and prediction. Neural representations explain how animals are able to think, recall, and perform inferences about a stimulus in the absence of that stimulus. Neural representations also explain the systematic dependence between on-line and off-line cognition: intervening on one systematically intervenes on the other in a way that is sensitive to representational content. Similarly, a visitor to London can think about, and modify, her map of the Underground in her hotel room, and this will affect, in systematic ways, how that map guides her future behaviour.

Neural representations are not the only tool to explain behaviour and cognition. In some contexts, alternative forms of explanation are preferable. However, neural representations play a major role in cognitive science; they animate our best accounts of our cognitive capacities. Neural repre-

sentations feature in the *description* of cognitive mechanisms, in the specification of the *causes* of thought and behaviour, in the *explanation* of behaviour and thought, and in *prediction* and *intervention* concerning the cognitive life of animals.[4]

Neural representations are ascribed to animals at almost every stage in their cognitive processing from low-level sensory and motor processing to high-level planning and inference. One of the most famous discoveries about the mammalian visual system is selectivity of response of cells in the early visual cortex (Hubel and Wiesel 1962). Certain neurons respond to certain characteristic stimuli (bars of particular orientations) more strongly than others. This suggests that neural activity carries information about, and is used by the brain to represent, some features of the world (e.g., lines and edges). Recent work on the visual cortex has focused on determining the nature of these representations, which stimuli various neurons are responsive to, how their response is optimised to efficiently represent natural environments, and the sensitivity of their response to top-down influences (Pasupathy and Connor 2001; Simoncelli and Olshausen 2001; Chirimuuta and Gold 2009). Similar kinds of response-selectivity has been observed in the primary auditory cortex (Schriener, Read, and Sutter 2000).

Cells in the inferior temporal cortex are selectively responsive to more complex environmental categories (faces, hands, and human bodies) in a way that is invariant to changes in stimulus size, contrast, colour, and exact location on the retina (Logothetis and Sheinberg 1996; Kanwisher, McDermott, and Chun 1997.) Some cells appear to be highly specific in their responses: tuned to particular emotional expressions, direction of eye gaze, or particular people (Perrett et al. 1985; Quian Quiroga et al. 2005). Their selective response appears to play a role in visual categorisation, learning, and memory (Milner and Goodale 2006). Activity in these areas is shared between on-line interactions with environmental stimuli and off-line experience such as visual imagery and hallucination (Albright 2012).

Neural representations play a major role in understanding memory and learning. One aspect of memory and learning that has received particular attention is spatial learning. Some neurons in the rat hippocampus are selectively responsive to spatial locations in certain environments (place cells), others are selectively responsive to head direction (direction cells), others in the primate hippocampus are selectively responsive to particular regions

of space falling into the field of view (spatial-view cells) (Moser, Kropff, and Moser 2008). Learning and memory are explained by long-term storage of, and associations between, these neural representations (Eichenbaum 2004).

The prefrontal cortex seems to be the primary neural basis of working memory; working memary holds an object briefly 'in mind' when it is no longer visible. The prefrontal cortex is important for decision-making tasks when, for example, animals have to make a decision about an absent stimulus, or a stimulus that has not yet occurred (Miller, Erickson, and Desimone 1996). Neural representations are the standard way to understand how this decision-making works: working memory involves the manipulation of neural representations, which can be present even if the stimuli they represent are absent (Miller and Cohen 2001).

Explanations of how humans understand each other often posit representations of the self, other humans, and belief and desire-like states of those humans. Recent attention has focused on identifying the neural basis of these representations. One influential suggestion is that they are based, in part, on the representations afforded by mirror neurons—neurons which fire when an animal acts and when the animal observes the same action being performed by another animal (Gallese 2007). Neural representations relevant to social cognition appear to be located in the medial frontal cortex (Amodio and Frith 2006).

Neural representations are also involved in motor activity and motor learning. Theories of these domains posit two types of neural representations: *forward models* are representations of the body that predict the sensory consequences of a given action allowing the animal to form quick anticipatory responses and cancel self-generated sensory signals; *inverse models* represent the body in a different way, allowing the animal to infer which motor commands to send to achieve a desired bodily position from its current state (Miall and Wolpert 1996). Flexible, rapid, and robust motor activity is explained by forward and inverse models working in tandem (Wolpert and Kawato 1998). The neural location of forward and inverse motor models is not yet established, but both are suspected to lie in the cerebellum (Wolpert, Miall, and Kawato 1998; Cerminara, Apps, and Marple-Horvat 2009).

4. Neural Representation Fictionalism (NRF)

Neural representations feature heavily in cognitive science. We saw in Section 2 that if our best description of the world contains ineliminable

reference to *X*s, then we are committed to the existence of *X*s. Cognitive science appears to be committed to the existence of neural representations.

This raises a problem. How do neural representations fit into the physical world? What elevates certain neural states to be *representations*?

Two tempting answers have to be avoided. First, it cannot be our *social conventions* that make a neural state a representation. Social conventions appear to be why many familiar public external representations —written language in English, signs, diagrams, the London Underground map—are representations. These count as representations because we adopt the social convention that certain marks on the page represent certain states of affairs. However, this cannot be true of neural representations. No appropriate social conventions are in play for neural representations to elevate them to representational status. Often we do not know *which* neural states represent, or *what* they represent. It is understood as cognitive science's job to discover the relevant neural representations, not to stipulate representational conventions, or search for hidden conventions in the social domain.

Second, it cannot merely be the *response-selectivity* of a neural state to a given stimulus that makes it a representation of that stimulus. As noted in Section 3, response-selectivity is often used to justify attribution of a neural representation. However, response-selectivity cannot be what makes a physical state a representation. Many physical states have response-selectivity but are not representations, and representations may occur without reliable response-selectivity (Ramsey 2007).

Neural representations must gain their representational status in a different way from conventional representations, and they cannot gain it from response-selectivity alone. Neural representations are claimed to achieve this feat by being *natural* representations. Natural representations gain their representational status independently of, and prior to, our social conventions. Since the 1970s, a great deal of attention has focused on trying to give a theory of natural representation. Accounts of natural representation often start with a simple response-selectivity condition and supplement or modify it with extra conditions in an effort to overcome its problems.[5] Unfortunately, and despite a large investment of effort, an adequate theory of natural representation not been forthcoming. Many contemporary philosophers suspect that representation simply cannot be naturalised.[6]

There appear to be two options. First, hard-headed Realism. We can continue to assume that neural-representation talk is, as it appears to be,

true, fact stating, and an ineliminable part of cognitive science. We accept that this entails a commitment to the existence of neural representations. However, we adopt an optimistic attitude and assume that the project of naturalising neural representations will eventually succeed. No account has succeeded so far, but perhaps an adequate account will be found. Second, Eliminativism. We excise neural-representation talk from serious fact-stating discourse in cognitive science. This may involve junking the entire approach that uses neural representations to explain cognition described in Section 3 (Beer 1995). Or, it may involve preserving neural-representation talk but quarantining it as an 'informal gloss' that can be safely paraphrased away when we ascend to the level of serious fact-stating talk (Chomsky 1995; Egan 2003). The cost of Eliminativism is that it requires painful revision to existing practice in cognitive science. Talking about neural representations is extremely useful; it is hard to eliminate, or paraphrase it away, and still achieve our ends.

Realism and Eliminativism are the main roads traveled. Realism lumbers us with the task of naturalising representation. Eliminativism requires painful revision to cognitive science. In this section, I propose a third option: Neural Representation Fictionalism.

What is Neural Representation Fictionalism (NRF)? According to NRF, neural-representation talk is false but serves an important purpose and, for that reason, should be preserved. NRF claims that neural-representation talk in cognitive science is perfectly in order and cannot, and should not, be eliminated or paraphrased away from serious fact-stating language. However, neural-representation talk does not bring with it any commitment to the existence of neural representations since it is understood as systematically false. Talking about neural representations is a useful device for cognitive science, but no more ontologically committing than talking about water as a *continuous incompressible fluid* is in fluid dynamics. This distinguishes NRF from Realism. What distinguishes NRF from Eliminativism is that a fictionalist interpretation of neural-representation talk is claimed to yield similar goods for cognitive science as Realism—explanatory, descriptive, causal, and instrumental goods. The intention of NRF is to allow us to reap the benefits of Realism without Realism's ontological costs.

According to NRF, statements of the following form are false:

(5) Neuron/brain region activity *X* represents *Y*.

However, statements of form (5) nevertheless serve a useful purpose and are fact stating. NRF is likely to endorse statements that are related to (5) but which concern a fiction. Precisely how to state these claims depends on the version of NRF being employed (see below). NRF is likely, however, to endorse something like the following:

(6) In the Neural Representation Fiction, neuron/brain region activity *X* represents *Y*.

Where the *Neural Representation Fiction* is the practice that attributes neural representations to the brain. According to NRF, the following claims are false:

(a) Some neural activity in V1 represents edges and lines.

(b) Some neural activity in the fusiform gyrus area represents faces.

(c) Some neural activity in the hippocampus represents spatial location and head direction.

(d) Some neural activity in the prefrontal cortex represents an absent stimulus in working memory.

But the following are true:

(a*) In the Neural Representation Fiction, some neural activity in V1 represents edges and lines.

(b*) In the Neural Representation Fiction, some neural activity in the fusiform gyrus represents faces.

(c*) In the Neural Representation Fiction, some neural activity in the hippocampus represents spatial location and head direction.

(d*) In the Neural Representation Fiction, some neural activity in the prefrontal cortex represents an absent stimulus in working memory.

In familiar fiction, e.g., the Sherlock Holmes stories, there is a written body of text that supplies the fiction. This text determines what is, and isn't, true according to the fiction. For NRF, there is no such text.

Instead, the Neural Representation Fiction should be understood as the *best agreed theory* about what neural states represent. This theory is implicit in the practice of cognitive science. Within that practice, researchers judge that certain neural states represent certain stimuli. In some cases, there is agreement that certain neural states represent, and about their representational content. There is also agreement about high-level features of the practice such as the kind of evidence sufficient to justify ascription of neural representation. To the extent that there is any agreement in the practice of cognitive science about judgements concerning neural representation, that practice can be taken as an implicit theory, which I will call the *Neural Representation Fiction*.

The relevant practice is a work-in-progress that will change as cognitive science develops. For many neural states, current practice is silent about whether, or what, they represent. Where the practice of cognitive science does not afford a coherent or determinate judgement about whether a neural state is a representation with a particular content, we can say that the relevant judgement, according to the Neural Representation Fiction, is (at least for the moment) *undetermined*. Where there is agreement in the practice, we can say that there is a fact about neural representation according to the Neural Representation Fiction. Concerning *how* agreement is brought about, I suggest that NRF defer to cognitive science. Cognitive science has its own standards for reaching agreement on what is or isn't a neural representation. It is not NRF's role to codify cognitive science's standards, or attempt to impose standards from the outside. The Neural Representation Fiction is constituted by what best agreed practice in cognitive science takes the facts about neural representation to be, where what is meant by *best* is best according to the standards for systematising that practice internal to cognitive science.

Now that we have laid out the basic strategy of NRF, there are many options for how the details of the view could be developed. Following Yablo (2001), one might distinguish at least the following options:

> **Instrumentalism:** the speaker is not 'really' asserting anything about neural representations, only pretending to do so.
>
> **Meta-fictionalism:** the speaker is 'really' asserting that according to the Neural Representation Fiction, the neural representations are so-and-so.

Object-fictionalism: the speaker is 'really' asserting that the world is in a certain condition, namely, the condition it needs to be in to make it true in the Neural Representation Fiction that the neural representations are so-and-so.

Figuralism: the speaker is 'really' asserting that *something* is in a certain condition, but perhaps not the world; the neural representations are functioning as representational aids in a figurative description of something else (the *Y*s), where the *Y*s may themselves be representational aids invoked to help us describe still further objects.[7]

There is also a range of further Fictionalist options for developing NRF. For example, van Fraassen (1980) argues that scientific theories should be understood as aiming at acceptance rather than belief, where acceptance is an attitude that falls short of belief. Yablo (2006) and Hinckfuss (1993) endorse a pragmatics/semantics distinction and argue that Fictional contexts are those in which certain assumptions are pragmatically presupposed. Eklund (2005) proposes that in certain contexts we make claims but remain indifferent to some of the implications expressed, including existential implications.

In this paper, I do not wish to privilege any one of these options for developing NRF over the others. My intention is instead to scout the general terrain of NRF and to raise two problems that apply to *any* form of NRF.

5. Objections to NRF

The promise of NRF is to let cognitive science enjoy the benefits of neural-representation talk without the cost of ontological commitment to natural representations. Two obvious tests of adequacy of NRF are: (a) whether NRF really allows us to avoid the task of naturalising representation; (b) whether NRF can yield the same benefits for cognitive science as Realism. In this section, I argue that NRF faces two objections that seem to show that it cannot meet either adequacy condition.

5.1. NRF doesn't avoid the task of naturalising representation

NRF, like all forms of Fictionalism, presupposes that a *fiction* exists. In the case of NRF, this is the Neural Representation Fiction. The fictions

used by NRF and other forms of Fictionalism are, by their nature, *representations*: they represent the world as *thus and so*. These fictions must be representational in order for it to be true, according to the Fiction, that the world is thus and so; or, for us to *accept*, or pragmatically *endorse*, the state of the world according to the Fiction.

A view that adopted a Fictionalist stance towards *all* representation talk—call it Global Representation Fictionalism—would be incoherent. Suppose this form of Fictionalism claimed to avoid ontological commitment to any representations; all talk of representation should be understood as true only according to a fiction. An immediate problem is that this view cannot escape commitment to at least one representation: the fiction that describes the representational facts. If this representation were not to exist, the Fictionalism would not make any sense. A truly Global Representation Fictionalism is, therefore, not viable.

NRF is not Global Representation Fictionalism, but NRF is vulnerable to a related worry. As mentioned in Section 3, it is widely assumed that neural representations are more fundamental than, and ground, other representations. Neural representations ground, and are somehow responsible for, personal-level thoughts such as beliefs, desires, and intentions. Personal-level representations in turn ground conventional representations such as signs, maps, and public language (Grice 1957; Lewis 1969).

Assume for the moment that NRF is correct. If NRF is correct, then there are *no such things* as neural representations: neural-representation talk should be understood as describing entities that occur in fiction. But if neural representations do not exist, then the grounding claim must be false. Neural representations cannot play the role of grounding other representations if they do not exist. One cannot use a fictional entity to ground (constitute, realise, or otherwise bring into existence) a real entity. Personal-level thought and conventional representations would have to be grounded in some other way.

There appear to be two options for an advocate of NRF at this point.[8] First, *keep the grounding claim*. The way to do this appears to be to broaden the scope of NRF to bring other representations into the fictional domain. One could adopt a fictionalist stance towards personal-level intentional states and conventional representations as well as to neural representations. This would allow the grounding claim to come out as true concerning the entities in the fictional domain. Unfortunately, this strategy

quickly runs into difficulties. As we saw above, NRF is committed to the existence of at least one representation, the Neural Representation Fiction. Hence, the grounding claim must be false at least for this representation. If the Neural Representation Fiction exists, it cannot be grounded in nonexistent entities. But if this one fiction is real, it is bewildering how it can exist by itself. In Section 4, we said that the Neural Representation Fiction is made up of the thoughts, intentions, judgements, and beliefs of cognitive scientists. This account of the Neural Representation Fiction is not available on the current strategy; none of the above mentioned representations exist, so they cannot make up the Neural Representation Fiction. An advocate of NRF is left with the problem of explaining how and why the Neural Representation Fiction exists as *the* lone *sui generis* real representation. And to the extent that an advocate of NRF admits other representations in order to explain the Neural Representation Fiction, she is forced to give up the grounding claim for them.

This leads to the second option: *drop the grounding claim*. This would allow us to keep personal-level intentional states and conventional representations. If one wished, personal-level thoughts could still ground conventional representations. One could also allow the Neural Representation Fiction to be made up from personal-level intentional states and conventional representations. One would have to give up, however, the claim that personal-level intentional states and conventional representations are grounded by neural representations. This strategy generates a different problem for NRF. The claimed benefit of NRF was that it allowed us to avoid the problem of naturalising representation. The current strategy leaves us with the problem of explaining *how*, if the grounding claim is false, personal-level intentional states and conventional representations gain their representational status and content. The grounding claim was designed to answer this: it aimed to naturalise personal-level intentional states and conventional representations via neural representations. But on the current strategy, this answer is no longer available. NRF understood this way reintroduces, with full force, the problem of naturalising representation. The original problem is transformed from that of naturalising *neural* representations to naturalising the *personal-level* representations and *conventional* representations on which the Fictionalist reading of neural-representation talk depends. The upshot is that NRF does not allow us to avoid the task of naturalising representation.

How might an advocate of NRF respond? One option is to argue that personal-level intentional states or conventional representations will prove easier to naturalise than neural representations. It is an open question whether the naturalising project should start with neural representations, personal-level thoughts, or conventional representations. There is some recent work that appears to show that beginning with conventional representations, without taking a detour through neural representations, has some promise.[9] If this were to prove correct, then perhaps this objection to NRF can be deflected.

5.2. NRF doesn't serve cognitive science as well as Realism

Neural representations are used by cognitive science for *prediction*, *description*, *intervention*, *causation*, and *explanation*. A key test for NRF, as for any form of Fictionalism, is whether it delivers the goods; whether the Fictionalist construal of the discourse in question serves our interests just as well as Realism. NRF promises fewer ontological commitments while yielding the same benefits as Realism—can it deliver on this?

At least some of the roles of neural representation in cognitive science appear apt to be served by fictional neural representations just as well as by real representations. Fictional neural representations appear to be able to serve cognitive science's interests for *prediction*. Just because an entity is fictional does not bar it from being useful in generating predictions. Fictions are often used to generate true predictions. In electrostatics, one might assume the existence of fictional mirror charges for generating predictions about the behaviour of real bodies. Fictional neural representations also appear apt to serve cognitive science's interests for *description*. Descriptions need not be true in order to feature in our best theories. The kinetic theory of gases is one of our best descriptive theories even though the entities that it posits—hard, perfectly-elastic, billiard-ball-like, atoms—do not exist. Fictional neural representations also appear apt to serve cognitive science's interests for *intervention*. Interventions guided by fictions can be just as successful as those guided by truth. Fictitious forces (the Coriolis force, the centrifugal force, the Euler force) may enter into our deliberation alongside real forces when we intervene on Earth-bound dynamical systems. The fact that these forces are fictitious does not make our interventions any less successful.

Fulfilling the *causal* and *explanatory* roles of neural representations in cognitive science appears to pose a more serious challenge for NRF. Let us take these two roles in turn.

First, *causal* role. In cognitive science, neural representations are assumed to be *causes* of behaviour and cognitive activity. *Prima facie*, this role for neural representations appears to be incompatible with their fictional status. In order for something to be a cause, that entity must exist. Fictional entities cannot cause, only real entities cause. Therefore, NRF appears unable to accommodate at least this role of neural-representation talk in cognitive science.

This objection to NRF may be less worrying than it may first seem. An advocate of NRF can reply that talking as if neural representations have a causal role is useful, even if false. One benefit of this talk is that it provides us with a way of referring to the (real) *neural causes* of behaviour. According to NRF, attribution of representational properties to neural states is systematically false. Yet, attributing representational properties to neural states, even if false, provides a way of *labelling* neural states, and hence of keeping track of them. We can use these (fictional) labels as a way to refer to the underlying neural states just as if we had given the neural states proper names. So even if it is false that neural representations cause behaviour, it can still be useful to assert this because it allows us to express—using the handy set of fictional labels that NRF provides—true causal relationships between neural states and behaviour. This offers at least one strategy for reconciling NRF with cognitive science's ascription of causal roles to neural representations.[10]

Second, the *explanatory* role of neural representations. This is more difficult for NRF to accommodate. As described in Section 3, one of the primary functions of neural-representation talk is to explain patterns of success and failure associated with animal cognition. Neural-representation talk is assumed to provide the *best explanation* of many cognitive phenomena. But, if ascribing neural representations is the best explanation of those cognitive phenomena, then according to Inference to the Best Explanation (IBE), we should *believe that such ascriptions are true*. However, this is flatly incompatible with NRF.

This objection to NRF can be summarised as follows:

1) Our best explanation of certain cognitive phenomena involve appeal to neural representations.

2) We ought rationally to be believe that our best explanations are true (IBE).

3) Therefore, we ought rationally to believe in the existence of neural representations.

The objection reveals an incompatibility between NRF and two ideas that are central to cognitive science: (1) neural representations best explain many cognitive phenomena, and (2) Inference to the Best Explanation.

How can NRF get around this? There appear to be two options, both of which have serious costs.

The first option is *reject (1)*. On this option, one would *not* claim that ascribing neural representations is the best way of explaining the relevant cognitive phenomena. This could keep IBE (premise 2) intact. The downside is that this strategy cuts against the motivation for NRF. We saw in Section 3 that one of the primary roles ascribed to neural representations is as the best explanation of cognitive phenomena. If we reject this and keep IBE, then it is incumbent on us to show that a *better* explanation exists of the relevant phenomena that avoids appeal to neural representations. But this was precisely the challenge that stymied Eliminativism and which NRF claimed to avoid. If NRF requires us to find a better nonrepresentational cognitive science, then it is unclear what advantage NRF has over Eliminativism.

The second option is *reject (2)*. This could be done in a number of ways. One way is to downplay the importance of explanation in science in general, as does van Fraassen (1980; 1985). On this view, explanation plays a relatively minor role in scientific practice compared to that of prediction, description, and intervention; good explanation is not a particularly significant matter as far as our ontological commitments are concerned. Alternatively, one might *keep* the importance of explanation in cognitive science, but *break* the link between truth and best explanation. This option appears to fit with recent work on scientific models as explanatory fictions (Bokulich 2011; Bokulich 2012; Frigg 2010a; Frigg 2010b). On such a view, explanatory value remains important, but it can be provided by a fiction just as well as by truth. The cost of both strategies is that cognitive science has to reject IBE. This appears to be a heavy cost indeed. IBE is one of the primary inferential methods in cognitive science. Our knowledge of internal cognition is based almost entirely on what best explains the behavioural and neural data. If we are not justified in inferring the truth of the best explanation of this data, then we appear to know almost nothing in cognitive science. This seems too high a price to pay for NRF.

6. Conclusion

We have seen that NRF faces at least two objections. These objections concern (i) whether NRF avoids the task of naturalising representation, and (ii) whether NRF adequately serves the interests of cognitive science. Both objections affect NRF independently of the exact form of Fictionalism that NRF employs.

Concerning (i), the best option for NRF appears to be to argue that the task of naturalising representation is on firmer ground with conventional representations or personal-level thoughts than it is with neural representations. Concerning (ii), NRF faces an unpleasant choice: either do the Eliminativist's work for her, or jettison IBE from cognitive science. Neither option is palatable. Perhaps the best strategy for an advocate of NRF is to find reasons for rejecting IBE in the case of neural representations that do not apply to other areas of cognitive science where IBE is employed. However, at this stage it far from clear what those reasons could be.

This article is only a first step in the exploration of NRF. The objections above indicate that further work is needed to show that NRF can deliver on its promise to have benefits over Realism and Eliminativism about neural-representation talk.[11]

Mark Sprevak

University of Edinburgh

NOTES

1. For example, see van Gelder (1995), Brooks (1991), Keijzer (1998), Beer (1995).
2. For example, see Bechtel (1998), Clark and Toribio (1994), Grush (2003), Markman and Dietrich (2000), Ramsey (2007).
3. Quine (1960); Quine (1980).
4. See, for example, Nicolelis and Lebedev (2009) on using motor neural representations to drive prosthetic limbs.
5. See Dretske (1981); Dretske (1995); Fodor (1990); Millikan (1984).
6. Concerns have come from a wide range of sources, with some principled worries coming from Loewer (1997); Putnam (1981); Kripke (1982); Ramsey (2007).
7. Summary of the options taken from Eklund (2011).
8. A third option, Fictionalism about grounding claims, will not be discussed here. I assume that any claim involving grounding should be understood in a Realist vein.
9. See Skyrms (2010). However, see Cao (2012) and Godfrey-Smith (2012) for an argument that Skyrms's view can be turned to the service of naturalising neural representation.

10. Note that this indicates a potential strength of NRF with respect to Realism about neural-representation talk: NRF allows us to avoid Kim's (1998) exclusion problem for representational properties. There is no threat that neural and representational causes systematically overdetermine behaviour, because talk of representational causes is really just a means of expressing a truth about neural causes.

11. I would like to thank an anonymous referee for helpful comments on a previous draft of this paper, and Tamás Demeter for his encouragement and wonderful work in editing this volume.

References

Albright, T.D. 2012. "On the Perception of Probable Things: Neural Substrates of Associative Memory, Imagery, and Perception," *Neuron* 74: 227–45.

Amodio, D.M., and C.D. Frith 2006. "Meeting of Minds: The Medial Frontal Cortex and Social Cognition," *Nature Reviews Neuroscience* 7: 268–77.

Bechtel, W. 1998. "Representations and Cognitive Explanations: Assessing the Dynamicist's Challenge in Cognitive Science," *Cognitive Science* 22: 295–318.

Beer, R.D. 1995. "A Dynamical Systems Perspective on Agent-Environment Interaction," *Artificial Intelligence* 72: 173–215.

Bokulich, A. 2011. "How Scientific Models Can Explain," *Synthese* 180: 33–45.

———. 2012. "Distinguishing Explanatory from Nonexplanatory Fictions," *Philosophy of Science* 79: 725–37.

Brooks, R.A. 1991. "Intelligence Without Representation," *Artificial Intelligence* 47: 139–59.

Cao, R. 2012. "A Teleosemantic Approach to Information in the Brain," *Biology and Philosophy* 27: 49–71.

Cerminara, N.L., R. Apps, and D.E. Marple-Horvat 2009. "An Internal Model of a Moving Visual Target in the Lateral Cerebellum," *Journal of Physiology* 587: 429–42.

Chirimuuta, M., and I. Gold 2009. "The Embedded Neuron, the Enactive Field?" in *The Oxford Handbook of Philosophy and Neuroscience*, ed. J. Bickle. Oxford: Oxford University Press.

Chomsky, N. 1995. "Language and Nature," *Mind* 104: 1–61.

Clark, A., and J. Toribio 1994. "Doing Without Representing?" *Synthese* 101: 401–31.

Demeter, T. 2009a. "Folk Psychology Is Not a Metarepresentational Device," *European Journal of Analytic Philosophy* 5: 13–38.

———. 2009b. "Two Kinds of Mental Realism," *Journal for General Philosophy of Science* 40: 59–71.

———. 2010. "In Defence of Empty Realism," *Journal for General Philosophy of Science* 40: 195–97.

Dennett, D.C. 1991. *Consciousness Explained*, Boston, MA: Little, Brown and Company.

———. 2002. "Quining Qualia," in *Philosophy of Mind: Classical and Contemporary Readings*, ed. D.J. Chalmers, Oxford: Oxford University Press.

Dretske, F. 1981. *Knowledge and the Flow of Information*, Cambridge, MA: MIT Press.

———. 1995. *Naturalizing the Mind*, Cambridge, MA: MIT Press.

Egan, F. 2003. "Naturalistic Inquiry: Where Does Mental Representation Fit in?" In *Chomsky and his Critics*, ed. L.M. Antony and N. Hornstein. Oxford: Blackwell.

Eichenbaum, H. 2004. "Hippocampus: Cognitive Processes and Neural Representations that Underlie Declarative Memory," *Neuron* 44: 109–120.
Eklund, M. 2005. "Fiction, Indifference, and Ontology," *Philosophy and Phenomenological Research* 71: 557–79.
———. 2011. "Fictionalism," in *The Stanford Encyclopedia of Philosophy*, ed. E.N. Zalta. 2011st ed., plato.stanford.edu/archives/fall2011/entries/fictionalism/.
Fodor, J.A. 1990. *A Theory of Content and Other Essays*, Cambridge, MA: MIT Press.
van Fraassen, B.C. 1980. *The Scientific Image*, Oxford: Oxford University Press.
———. 1985. "Empiricism in the Philosophy of Science," in *Images of Science*, ed. P.M. Churchland and C. Hooker, Chicago: University of Chicago Press. 245–308.
Frigg, R. 2010a. "Models and Fiction," *Synthese* 172: 251–68.
———. 2010b. "Fiction and Scientific Representation," in *Beyond Mimesis and Conventions: Representation in Art and Science*, ed. R. Frigg and M. Hunter. Dordrecht: Springer.
Gallese, V. 2007. "Before and Below 'Theory of Mind:' Embodied Simulation and the Neural Correlates of Social Cognition," *Philosophical Transactions of the Royal Society of London, Series B* 362: 659–69.
van Gelder, T. 1995. "What Might Cognition Be, If Not Computation?" *The Journal of Philosophy* 91: 345–81.
Godfrey-Smith, P. 2012. "Signals, Icons, and Beliefs," in *Millikan and Her Critics*, ed. D. Ryder, J. Kingsbury, and K. Williford, Oxford: Wiley-Blackwell.
Grice, P. 1957. "Meaning," *Philosophical Review* 66: 377–88.
Grush, R. 2003. "In Defense of Some 'Cartesian' Assumptions Concerning the Brain and its Operation," *Biology and Philosophy* 18: 53–93.
Hinckfuss, I. 1993. "Suppositions, Presuppositions, and Ontology," *Canadian Journal of Philosophy* 23: 595–618.
Hubel, D.H., and T.N. Wiesel 1962. "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *Journal of Physiology* 160: 106–54.
Kanwisher, N., J. McDermott, and M.M. Chun 1997. "The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception," *Journal of Neuroscience* 17: 4302–11.
Keijzer, F.A. 1998. "Doing Without Representations Which Specify What to Do," *Philosophical Psychology* 11: 269–302.
Kim, J. 1998. *Mind in a Physical World*, Cambridge, MA: MIT Press.
Kripke, S.A. 1982. *Wittgenstein on Rules and Private Language*, Cambridge, MA: MIT Press.
Lewis, D.K. 1969. *Convention*. Cambridge, MA: Harvard University Press.
Loewer, B.M. 1997. "A Guide to Naturalizing Semantics," in *A Guide to Naturalizing Semantics,* ed. B. Hale and C. Wright, Oxford: Blackwell, 108–26.
Logothetis, N.K. and D.L. Sheinberg 1996. "Visual Object Recognition," *Annual Review of Neuroscience* 19: 577–621.
Markman, A.B. and E. Dietrich 2000. "In Defense of Representation," *Cognitive Psychology* 40: 138–71.
Miall, R.C. and D.M. Wolpert 1996. "Forward Models for Physiological Motor Control," *Neural Networks* 9: 1265–79.
Miller, E.K., C.A. Erickson, and R. Desimone 1996. "Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque," *Journal of Neuroscience* 16: 5154–67.

Miller, E.K., and J.D. Cohen 2001. "An Integrative Theory of Prefrontal Cortex Function," *Annual Review of Neuroscience* 24: 167–202.
Millikan, R.G. 1984. *Language, Thought and Other Biological Categories*, Cambridge, MA: MIT Press.
Milner, D., and M. Goodale 2006. *The Visual Brain in Action*, 2nd ed., Oxford: Oxford University Press.
Moser, E.I., E. Kropff, and M.-B. Moser 2008. "Place Cells, Grid Cells, and the Brain's Spatial Representation System," *Annual Review of Neuroscience* 31: 69–89.
Nicolelis, M.A. and M. A. Lebedev 2009. "Principles of Neural Ensemble Physiology Underlying the Operation of Brain-Machine Interfaces," *Nature Reviews Neuroscience* 10: 530–40.
Nolan, D., G. Restall, and C. West 2005. "Moral Fictionalism Versus the Rest," *Australasian Journal of Philosophy* 83: 307–30.
Pasupathy, A., and C.E. Connor 2001. "Shape Representation in Area V4: Position-Specific Tuning for Boundary Conformation," *Journal of Neurophysiology* 86: 2505–19.
Perrett, D.I., P.A.J. Smith, D.D. Potter, A.J. Mistlin, A.D. Milner, and M.A. Jeeves 1985. "Visual Cells in the Temporal Cortex Sensitive to Face View and Gaze Direction," *Proceedings of the Royal Society, Series B* 223: 293–317.
Putnam, H. 1981. *Reason, Truth and History*, Cambridge: Cambridge University Press.
Quian Quiroga, R., L. Reddy, G. Kreiman, C. Koch, and I. Fried 2005. "Invariant Visual Representation by Single Neurons in the Human Brain," *Nature* 435: 1102–07.
Quine, W.V.O. 1960. *Word and Object*, Cambridge, MA: MIT Press.
———. 1980. "On What There Is," in *From a Logical Point of View*, Cambridge, MA: Harvard University Press, 1–19.
Ramsey, W.M. 2007. *Representation Reconsidered*, Cambridge: Cambridge University Press.
Schriener, C.E., H.L. Read, and M.L. Sutter 2000. "Modular Organization of Frequency Integratation in Primary Auditory Cortex," *Annual Review of Neuroscience* 23: 501–29.
Simoncelli, E.P., and B.A. Olshausen 2001. "Natural Image Statistics and Neural Representation," *Annual Review of Neuroscience* 24: 1193–216.
Skyrms, B. 2010. *Signals*, Oxford: Oxford University Press.
Wallace, M. 2007. "Mental fictionalism," www.unc.edu/~megw/MentFic.pdf.
Wolpert, D.M., R.C. Miall, and M. Kawato 1998. "Internal Models in the Cerebellum," *Trends in Cognitive Sciences* 2: 338–47.
Wolpert, D.M., and M. Kawato 1998. "Multiple Paired Forward and Inverse Models for Motor Control," *Neural Networks* 11: 1317–29.
Yablo, S. 1998. "Does Ontology Rest on a Mistake?" *Proceedings of the Aristotelian Society, Supplementary Volume* 72: 229–61.
———. 2001. "Go figure: A Path Through Fictionalism," *Midwest Studies in Philosophy* 25: 72–102.
———. 2006. "Non-Catastrophic Presupposition Failure," in *Content and Modality: Themes from the Philosophy of Robert Stalnaker*, ed. J. Thomson and A. Byrne, Oxford: Oxford University Press, 164–90.