Contents, Vehicles, and Complex Data Analysis in Neuroscience

Daniel C. Burnston

Forthcoming in *Synthese*.
Penultimate draft. Please cite final version.

Abstract

The notion of representation in neuroscience has largely been predicated on localizing the components of computational processes that explain cognitive function. On this view, which I call "algorithmic homuncularism," individual, spatially and temporally distinct parts of the brain serve as vehicles for distinct contents, and the causal relationships between them implement the transformations specified by an algorithm. This view has a widespread influence in philosophy and cognitive neuroscience, and has recently been ably articulated and defended by Shea (2018). Still, I am skeptical about algorithmic homuncularism, and I argue against it by focusing on recent methods for complex data analysis in systems neuroscience. I claim that analyses based on machine learning tools, such as principle components analysis and linear discriminant analysis, prevent individuating vehicles as algorithmic homuncularism recommends. Rather, each individual part contributes to a global state space, trajectories of which vary with important task parameters. I argue that, while homuncularism is false, this view still supports a kind of "vehicle realism," and I apply this view to debates about the explanatory role of representation.

1. Introduction

Debate about representation in cognitive science is ultimately a discussion about explanation. In virtue of what does positing representations in a system like the brain explain behavior? Shea (2018) has recently given an account which attempts to ground representation in what I will call "algorithmic homuncularism" (AH). On this kind of view, distinct physical parts of a system serve as vehicles for distinct contents, and the causal interactions between those vehicles implement the content transformations called for by an algorithm.

Shea's account is exemplary of a widespread way of thinking about representation in the cognitive and neurosciences. However, I will argue that there are strong reasons to question AH. In particular, certain forms of complexity in neural population responses prevent assigning different contents to spatially and temporally distinct parts of the system.

I will make the point through close examination of data analysis methods as employed in primate electrophysiology. In massively multifunctional parts of the brain like the prefrontal cortex, groups of cells do not clearly divide up in terms of the task parameters for which they are selective. Rather, there is a consistent admixture of selectivity at the individual cell level. To make sense of the function of these populations, researchers have turned to methods for analyzing complex multivariate data, including *principal components analysis* and *linear discriminant analysis*. These techniques reveal patterns in population activity via dimensionality reduction.

I will argue that the particular explanations offered by these frameworks do not support dividing brain systems as AH proposes. The basic issue is that, rather than being spatially or temporally divided into distinct content-bearing parts, the explanatory properties of the system are patterns of whole population activity. I then consider what this means for realism about representation. I suggest that Shea's "vehicle realism" – the idea that there must be distinct vehicles of content – is on the right track, but that this does not require AH. Rather, we can individuate vehicles in terms of the distinguishable influence that different task parameters have on their dynamics. This view, I contend, supports (at least) a limited realism about neural representation.

I proceed as follows. In section 2 I lay out some of the broader issues surrounding the current discussion. In section 3, I articulate Shea's view and its commitments. In section 4, I introduce the data analysis techniques, and two case studies employing them. In section 5 I argue that the results of these investigations cannot be accounted for by Shea's view. In section 6 I advocate an alternative form of realism not based on AH. Section 7 concludes.

2. Setup

While in the remaining sections I will focus primarily on Shea's account, I do so because it is exemplary of a widespread kind of view. Here I want to point to the larger picture and clarify some of the surrounding terrain.

The idea I wish to isolate and criticize, which I will call "algorithmic homuncularism" (AH), is bound up with a certain view of reductive explanation in the neurosciences. It rests on a particular way of syncing up the notions of computation, representational content, and representational vehicle. On AH, explanation consists in (i) identifying algorithms that perform cognitive functions, and (ii) isolating the contents and transitions posited by those algorithms in distinct parts of the brain. The ideal explanation, on AH, determines precisely what computations constitute cognitive processes, and links these in a one-one manner with vehicles and causal transitions in the brain.

An immediate caveat is in order. Algorithms are mathematical, "medium independent" (Elber-Dorozko & Shagrir, 2019) posits, and themselves are not individuated in terms of contents, which generally involve referential relations to the environment (Chirimuuta, 2017; Egan, 2010, 2014a, 2014b). As such, AH is often conjoined with an attempt to naturalize content in one of the ways that philosophers have traditionally analyzed – e.g., informational or teleological views.

AH inherits much of the appeal of computer functionalism in philosophy of psychology. A standard view of how psychological functions are implemented is that the structure of the function is mapped to the causal structure of what implements it (Cummins, 1985, 2000; Fodor, 1975). This gets one the standardly appealing picture of explanatory levels, realism about psychological explanation, and so on. Moreover, at least according to some, it provides a way to link computational and mechanistic explanation (Elber-Dorozko & Shagrir, 2019; Piccinini & Craver, 2011).[1] Whereas the computational description of the system is mathematical, it provides explanatory targets – i.e., isolating vehicles and causal relations – for neuroscientific investigation (Povich, 2015, 2019). As such, AH hopes to align philosophical theorizing with neuroscientific practice, which sounds like a good thing.

Moreover, AH brings vital insights on board about when and why we resort to representation talk. First, we posit contents to understand particular kinds of functional relationships between organisms and environments. Second, representations are used to understand the *organization of the system* that implements those relationships. As Rupert (2018) puts it, representational posits are valid when they "play an explanatory role in an architecture" (p. 205) implementing cognitive processes, including the neural architecture (Rathkopf, 2017).

Unfortunately, I think there are strong reasons to doubt AH. To see why, let's consider its commitments in slightly more detail. AH is committed to the conjunction of what I will call *injective mapping* and *causal isomorphism*. Injective mapping is the view that the contents processed in the algorithm are realized in spatially and functionally distinct vehicles. This means that for each stage posited in the algorithm, there will be a particular physical part that realizes the postulated content, and this part will be different from the vehicles for other stages. Causal isomorphism implies that for each computational step in the algorithm, there will be a particular causal relationship between those distinct vehicles. The causal steps "accord" with the algorithm, and "process" the contents posited in accordance with the transformation the algorithm describes. Such a process, on which different contents are generated in stage-based processing, is a natural way of glossing the idea that representations are "used" or "consumed" within the system (Bechtel, 2016).[2]

[1] For skepticism regarding this kind of mapping, see (Chirimuuta, 2014; Weiskopf, 2015).

[2] This is a weaker notion of consumption that Millikan's (1989) classic "consumer semantics," since on this view one needn't *individuate* contents in terms of their consumers. Shea (2018) rejects the traditional

AH is widely held by both philosophers and neuroscientists. Here are just a couple of recent examples, one from a philosophical and one from a neuroscientific context, which I take to be indicative of a much wider trend.

- Many of [neural systems'] components (columns, nuclei) … contain representations and perform more limited computations over them; the computations they perform are component processes of the computations performed by their containing systems. Therefore, large neural components are representational and computational, and the same holds for their components (e.g. networks and circuits). Again, the computations performed by smaller components are constituents of the larger computations performed by their containing systems, and that is how the computations of their containing systems are mechanistically explained. (Boone & Piccinini, pp. 1523-1524)

- "Because local circuits perform highly specialized computations or process information from different sources, distinct behavioural tasks require different combinations of regions to work together, calling for different patterns of information flow through long-range anatomical connections" (Akam & Kullman, 2014, p. 111)

Despite its wide appeal and theoretical advantages, I will argue that AH is false, at least if certain trends in theorizing in systems neuroscience are on the right track. In particular, I contend that certain kinds of complexity in neural functioning prevent describing neural systems as AH recommends. To get to the argument, I will outline several approaches to population analysis in electrophysiology. The most important aspect of these approaches for current purposes is that, rather than trying to isolate distinct contents to temporally and spatially distinct parts of a system, they try to show how environmental and task information is represented in patterns of activity across an entire population.[3]
Put differently, representations are realized as trajectories in the state space of a whole population, rather than as isolated to distinct parts of that population. Hence, injective mapping is false.

consumer semantics approach, opting for a combined teleofunctional and causal account of how vehicles get their contents. On this kind of position, one needn't posit that representations at one vehicle are "read" by later vehicles (Shea also rejects this notion of consumption). The sense of 'use' at work here is underlain by there being vehicles that are causally related so as to implement the steps of the algorithm.

[3] Throughout this paper, I will use an intuitive notion of "pattern." I have elsewhere defined a pattern as a "type of quantitative variation," (Burnston, 2017c) and I believe this definition covers my use here (cf. Kästner & Haueis, 2019). For a full discussion of the discovery of patterns via PCA and LDA, see Millhouse (forthcoming).

If these analyses genuinely describe the functioning of the population, then the causal isomorphism commitment is also undermined, because there is no sense in which causal interactions between distinct content-bearing vehicles mirror those posited in an algorithm (cf. Cao, 2011). The approach I prefer is to analyze representation in terms of the underlying dynamics of the system. This approach is present in the literature, but is rarely explicitly argued against by proponents of AH.[4] Of course, there is considerable debate about whether explanations focusing on dynamics are compatible with the notion of representation (Bechtel, 1998; Chemero & Silberstein, 2008).[5] The contributions of this paper are: (i) to argue that data analytic methods suggest pursuing a dynamic approach that eschews injection mapping and causal isomorphism, and thus AH, and (ii) that this approach is compatible with (a variety of) representational realism.

My strategy will be to focus on one of Shea's core commitments, namely "vehicle realism." Shea argues – rightly in my view – that in order for representational explanation to be feasible, it must cite representations in describing the functional organization of the system. Hence, there must be physical vehicles that are realizers of contents. As we will see in the next section, Shea does not distinguish this view from AH. However, I claim that one can have vehicle realism without AH. Populations, as analyzed in the methods to be described, are not organized according in the way suggested by AH. This does not mean that they do not represent the important features of environments and particular tasks. Moreover, I will suggest that vehicles of content can exhibit "algorithmic coherence" (AC) – correlations between elements of their activity and the contents posited in an algorithm – without AH being true.

In the next section, I will discuss Shea's particular version of AH. In section 4, I will introduce the data analysis techniques that inform the alternative view, and two case studies. Section 5 argues that AH misdescribes results produced by these analyses, and section 6 describes an alternative form of realism.

---

[4] Clark (1998) gives an early statement of the idea that understanding the dynamics of a system can be a key to understanding representation. The fully-fleshed out proposal that is closest to mine is by Shagrir (2012b), who also analyzes representation and computation in terms of transitions between locations in state space. The view I will express here is largely in league with his.

[5] There are a variety of other important issues in the surrounding terrain as well. The relationship between algorithms and realizers is at the heart of an interesting debate about how computations themselves are individuated (Piccinini, 2007; Shagrir, 2012a). Further, it is contested how we should mutually analyze the intersecting notions of information, representation, and computation (Piccinini, 2008; Piccinini & Scarantino, 2011). And the type and way in which the brain implements algorithms has ramifications for the discussion of whether computation in the brain is analog or digital (Maley, 2018; Piccinini & Bahar, 2012; Piccinini & Shagrir, 2014; Shagrir, 2006, 2010). I will not have space to address these issues here.

3. The Commitments of Representationalism?

In the remainder of the paper I will focus on Shea's (2018) recently developed account of representational realism. This is because I agree with many of Shea's commitments and much of his apparatus is helpful. In addition, he is explicit in his endorsement of, and clear in his phrasing of, AH. (I thus intend my focus on Shea as salutary.) There are three aspects of Shea's account that I endorse. The first is his commitment to *vehicle realism*, the second is his discussion of representational relations, and the third is his view of function. I'll go through these very briefly before getting to his articulation of AH.

Vehicle realism is the view that there are physical parts and processes of a representational system to which contents can justifiably be assigned. "Parts" here should be construed loosely as including cells, groups of cells, and brain areas. The processes that carry contents in the brain are electrical (and chemical), but this too should be construed somewhat loosely. Individual spike trains, population codes, and local field potentials have all been posited as potential carriers of content. The appeal of vehicle realism is that it ties content explanation to causal explanation, and thus captures the intuition that explanations describe causal processes in the system of interest. Fitting representation into causal explanation is one way to substantiate its explanatory value, although it raises worries that the content claim does not contribute anything over and above a pure causal story (Rescorla, 2014). I will discuss this further in section 6.

According to Shea, content is based on the establishment of "exploitable relations" to the environment. A vehicle gets its content through bearing some relation to the external world, but there are multiple kinds of relation that will do the trick, including simple correlation and structural mapping. I agree that the brain has multiple ways of representing information, and hence think there is good reason to be wide-net about content-determining relations in this way.

Lastly, Shea has an appealingly flexible notion of "task function," which captures much of how tasks are thought about in neuroscience. Shea argues that a notion of function is necessary for representational explanation to be of use, but thinks that functions can be determined in a variety of ways, including by natural selection, by intentional selection, and by learning history. What matters is that a structure be developed that allows the system to produce stable and robust beneficial outcomes in the relevant circumstances. Even on-the-fly learning can establish these functions – an account that Shea refers to as the "very modern history" (2018, p. 63) etiological account. Again, I agree that the structure of the task heavily influences what and how the brain represents (Burnston, 2016, forthcoming).

So, I will take all of these aspects of Shea's analysis as read. I do not think that any of this, however, requires endorsing AH. Shea, for his part, does not seem to distinguish vehicle realism from AH. Consider the following:

- Representation arises where a system implements an algorithm for performing task functions. That in turn has two aspects: internal vehicles stand in exploitable relations to features of the environment which are relevant to performing the task; and processing occurs over those vehicles internally in a way that is appropriate given those relational properties. (Shea, 2018, pp. 75-76)

- Internal processing implements transitions between vehicles … transitions called for by an algorithm which is suited to producing the input–output mapping given by the system's task functions. (Shea, 2018, p. 76)

- To count as explanatory, an algorithm will generally have different contents at different stages. The computation of what to do is mediated through a complex series of internal states. (Shea, 2018, p. 86)

On the position expressed here, a representational explanation of a robust outcome is feasible if spatially and temporally distinct vehicles carry distinct contents, and if the causal interactions between them capture the content transformations posited in an algorithm.

Hence, Shea endorses AH. One strong motivation Shea gives for the view is that there must be a *fact of the matter* about the way the organism represents the world. Given some stimuli and a robust outcome, there are innumerably many potential processes that could produce the behavior. If we are to be realists about representation, then we must find out which out of those possible accounts is the right one. In keeping with AH, Shea suggests that the way to do this is to find individual parts of the system that bear particular exploitable relations to the environment, and causal relations between them that implement the sequence of steps suggested in the algorithm. Shea thus claims that his view is a version of the kind of computational functionalism I discussed in the last section.

In the next section, I introduce the data analytical methods and results that I take to cause trouble for AH. Importantly, Shea specifically considers one of these analyses, and attempts to describe it in terms of his account. Hence, these projects fall under his intended scope.

4. Data Analytic Methods for Understanding Neural Populations

4.1. Methods

The standard approach to determining the information represented in a neuron or population is to assess its "selectivity." In this method, changes in neural firing rate are correlated with changes in experimental variables. Often, the variable that explains the most variance in cell firing is the one that the cell is taken to represent. This covers representation in the "correlational" sense, but populations of cells with different selectivities can be taken as a structured representation of a domain – consider, for instance, maps of orientation in V1 or of motion energy in MT.

So, one hope for AH would be to simply correlate given cells or populations with the relevant task parameters, and then describe interactions between them as implementing content transformations according to a posited algorithm. Indeed, these kinds of analyses are widely pursued. However, recent results raise challenges for this approach. There is widespread evidence that neurons exhibit *mixed selectivity* (Rigotti et al., 2013). This is to say that they are not selective for one experimental variable, but several. (So, in multiple regression analyses, several environmental and/or task variables non-redundantly explain the variance in the cells' responses.) These properties are widespread in perceptual and frontal cortices, as well as in subcortical areas (Panzeri, Macke, Gross, & Kayser, 2015; Saez, Rigotti, Ostojic, Fusi, & Salzman, 2015).

Partially to make sense of these and similar properties, complex data analysis methods have become part of the neuroscientist's basic tool kit. The basic conceptual framework for these tools is to construe one's data as constituting a high-dimensional space. Generally, each dimension is a measure of neural activity. So, it might be the spike rate recorded from a single cell, or the amplitude of oscillations of a certain frequency measured from an EEG electrode. A recording from a given trial will constitute a value along every dimension, and the dataset is the entire set of values recorded. Of course, this results in an extremely complex dataset. Analytic methods are used to reduce this dimensionality and say things in general about the population's responses.

There are two particular techniques I will discuss (cf. Millhouse, forthcoming). The first is called "principal components analysis" (PCA). PCA defines a small number of dimensions that capture the *variance* in the data. In PCA, the investigator begins by producing a covariation matrix, which measures the covariance between any two dimensions across the entire dataset. So, the first row of the matrix will be the first unit, and the columns in this row will be its covariance with every other unit, and so on.

The principal components are eigenvectors of this matrix. Leaving aside some of the linear algebra detail, the effect of the eigenvector manipulation is to construct a series of dimensions which capture the patterns of covariation. The "first" principal component captures the most variance in the dataset, the "second" principal component the second most, and so on. In principle, one could construct a number of components equal to the

dimensionality of the original dataset. But the point of the transformation is that usually the first handful of principal components will explain the majority of variance in the data. Put another way, any state of the data can be perspicuously described in terms of how much each component is contributing to that state. These are often called "loadings" or "weights" of the principal components. So, any state of the entire dataset can be explained by a few variables, namely the loadings on the principal components; the data is thus described in the "principal component space" or the "PC space."

PCA is an "unsupervised" data analysis tool. It simply attempts to describe the variance in the data as efficiently as possible. Other techniques are "supervised," which means that knowledge about the categories the data comprises is employed in constructing the analysis. Suppose that, in the high-dimensional dataset, we know that there are different categories, and which category each measured data point falls under. The goal is then to describe a small number of dimensions that best *separate* the categories in that space.

One supervised method of this kind is "linear discriminant analysis" (LDA). LDA takes as its goal constructing dimensions that will separate already known categories in the data. If the data had just two categories, for instance, LDA would attempt to draw a single dimension that maximized the distinction between them in the high-dimensional dataset. Any individual point in the space would then be "projected" onto that dimension, such that the maximum number of examples of category 1 possible would be on one end of the dimension, and the maximum number of examples of category 2 possible would be on the opposite end. As such, any new point from the original data space – or any subsequently measured point – could then be projected onto the dimension as a way of predicting the category to which it belonged. Three or more categories are distinguished by "hyperplanes" (Rich & Walls, 2016) through the space, such that each data point belonging to a category falls into one section cordoned off by the plane.

Analyses based on these tools have a certain structure, which I will argue is important for considering what they tell us about neural function. The lower-dimensional description of the data is *derived from the entire data set*. So, all measurements contribute to the lower-dimensional representation. One might think of this as a description of all of the states of the system under the conditions of measurement. Once the lower-dimensional description has been generated, however, one can *then* analyze *particular* states of the system (measurements) in terms of how they fit into the overall structure. So, we might analyze a neural population response in a particular trial in terms of how it compares to the overall set of states the population might be in. Further, we can even look *within* a trial to see how the population evolves in particular task conditions. These capacities have made dimensionality reduction techniques extremely important for understanding dynamic population behaviors in the brain. What is important about this, for current purposes, is that any measure of representation of a particular variable is taken to be constituted by the

whole population, rather than by distinct parts within that population. If so, then the system cannot be divided in the way AH recommends. I now pursue this argument in two particular cases.

4.2. Neuroeconomics

The field of "neuroeconomics" has recently tried to understand the processes by which organisms make value-based decisions. Intriguingly, cells in the orbitofrontal cortex (OFC) have shown selectivity for variables involved in choice situations – including the identity of the potential choices and their value (usually computed as subjective desirability minus effort required to obtain). Other cells have shown selectivity for *chosen value*. This is the absolute value of the object chosen, independently of its identity.

Very much in keeping with AH, models have been put forward on which value-based decisions are the result of a particular, ordered set of interactions between these cells. As noted by Padoa-Schioppa and Conen, this kind of stage-based approach is heavily influenced by economic thinking: "A core idea rooted in economic theory is that choosing entails two mental stages—values are first assigned to the available options and a decision is then made by comparing values" (Padoa-Schioppa & Conen, 2017, p. 736). So, the idea is that the options and values are represented, and then the chosen value – the quantity that reflects the choice – is computed at a subsequent stage. One such model is pictured below.

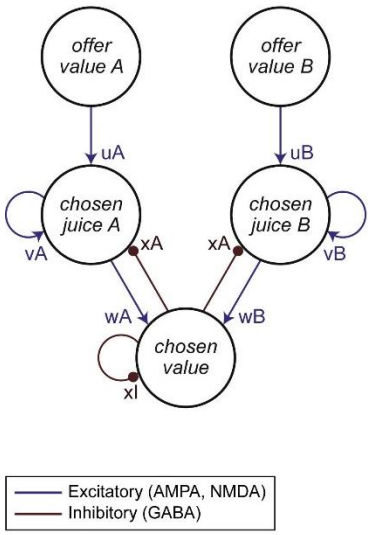

Figure 1. From Padoa-Schioppa and Conen (2017).

This view closely aligns with AH, in that it posits distinct representations at distinct stages, realized in different populations of cells. However, other investigations have cast this kind of explanation of the OFC into doubt. In particular, other authors have contended that the algorithmic approach mistakenly posits chosen value as a distinct, causally interacting

representation within the system, and that this sequential processing view fails to accurately capture the functional dynamics of the system. Both PCA and LDA have been employed to argue for this conclusion.

One way of looking at population activity is by measuring *local field potentials* (LFPs) – oscillatory patterns at the population level that reflect a sum of overall electrical activity in the population (Burnston, 2019; Canolty & Knight, 2010; Watrous, Fell, Ekstrom, & Axmacher, 2015). LFPs oscillate at particular frequencies and amplitudes, and changes in frequency and amplitude often correlate with changes in function. LFPs can be measured both intracranially through electrodes, and extracranially via MEG or EEG. Hunt et al. (Hunt, Behrens, Hosokawa, Wallis, & Kennerley, 2015) decided to check some of the commitments of the algorithmic approach by looking at how LFP patterns within trials interact with individual cell properties determined by more standard regression analyses.

The researchers had monkeys perform a cost/benefit task, on which they had to choose between two options, each of which had a value and a cost (e.g., a delay before reward). They recorded from individual cells and measured LFPs. Their use of PCA focused primarily on the LFP data. They assessed the first two principal components, and in particular what weightings on these components tended to correlate with across time in the population. Weightings on the first component tended to best correlate with LFP amplitude. The second component primarily accounted for the *temporal derivative* of the LFP signal. That is when weight on the second component was high, the LFP signal ramped up quickly during a trial, and when it was low it ramped up slowly.

There is a functional interpretation to this pattern, namely that on certain kinds of trials activity would tend to ramp up quickly, and thus be correlated with high weightings on PC2. Indeed, on *high value* trials, reaction times were quicker, LFP activity ramped up faster, and PC2 weights were higher. This suggests that PC2 reflected certain properties of the dynamics of the choice. The rub, however, came when the researchers used the PCA results to, in turn, analyze the single-cell properties. They showed that, when PC2 weights were used as a *co-regressor*, they explained away the effect of chosen value on cell responses. That is, the correlation with chosen value did not account for the cell's responses beyond the correlation with PC2 weight.

The AH-inspired approach suggests that cells representing chosen value receive inputs from distinct populations representing choice identity and choice value, and use those signals to compute the value that will drive the choice. However, the PCA measurements are derived from *the whole population* of cells (including those that correlate with chosen value). What the results suggest is that the chosen value signal is the outcome of how the entire population evolves, rather than occurring at a particular location or stage within that population. Hence, Hunt et al. suggest that the correlation with chosen value does not

underlie a representation that interacts causally with other representations but instead is a by-product of the temporal dynamics of the population, particularly the fact that population activity ramps up faster on high-value than low-value trials.

We can unpack this further by responding to an objection. One might suggest that, on a correlational view of representation, the fact that the correlation of cells to chosen value is accounted for by population dynamics does not undermine viewing them as representations – they could, for instance, be tracking chosen value through the temporal dynamics of the population.[6] Note, however, that this claim is insufficient to ground AH, which not only posits representations, but posits distinct causal roles for those representations interacting with other parts of the system. AH implies both *spatial* and *temporal* divisions between distinct stages of processing. What the PCA-results suggest is that chosen value is not an outcome of a sequence of processing stages that begin with other distinct populations and end in the choice, but rather that the whole system evolves temporally towards a choice.

Now, chosen value was the only property of selectivity that could be explained in this way. The representations of values associated with particular objects remained explanatory when co-regressed with either PC1 or PC2. As such, one might contend that a different algorithmic view might in fact capture the population responses.

There are reasons to question this too, however, which point to deeper functional principles in the system not revealed by the algorithmic view. In a subsequent study, Rich and Wallis (2016) used LDA to assess how values were represented dynamically in the OFC during choice. They trained monkeys on the reward values associated with four distinct pictures. Then, during each test trial, two of the pictures were shown, and the monkeys had to saccade to the picture whose associated reward they wished to receive. LDA was run to group places in the overall dataset according to picture values. This model could then in turn be analyzed to see which picture value was reflected in the population at any given time.

Intriguingly, the analysis showed that the population *switched back and forth* between reflecting the values of the two options during the course of a given trial. The model never suggested that the population encoded options that were unavailable, confirming that these signals were related to the choice process. Moreover, the *number* of switches reflected a variety of properties of choices, for instance the difference between the values of the options and the time the monkey took to choose. Fewer switches occurred when the values of the options were far apart, and more occurred when they were closer. Similarly, reaction times were greater when there were more switches within the populations.

[6] Thanks to an anonymous reviewer for pushing me to consider this objection.

Rich and Wallis suggest that the dynamic switching between option values thus reflects the process of *deliberation*. If what the population is doing is considering the value of each object in turn, then one would predict more switching when the objects are closer in value, and hence when the decision is harder. And these more difficult decisions should be reflected in longer reaction times. Note again the difference between this view of the dynamics and the AH view. AH suggests that the relative valuesm of the objects are encoded at a particular stage in an ongoing computational process. The LDA results, alternatively, suggest that this comparison occurs dynamically right up until the choice is made.

So, I suggest that these results tell against AH. They don't, however, tell against AC. As noted, corollaries of important decision variables can be uncovered both within the population and in individual cells at many distinct points. Hunt and Hayden (2017) suggest that "the algorithm is an emergent property of the system: certain correlates of value could therefore emerge naturally as a consequence of how neural dynamics unfold across different trials" (p. 173). Leaving aside complications about how to read "emergence" here (Burnston, 2019), we can note that the population's activity reflects important decision variables, without those variables being distinctly represented in a stage-based process. Moreover, the monkeys' behaviors can still be described as implementing certain value trade-offs, and since it is the dynamics of the choice system that produces this behavior, the process is of course coherent with those functional descriptions. None of this requires that AH obtain.

4.3. The Dynamics of Decision

A recent influential study from Mante et al. (Mante, Sussillo, Shenoy, & Newsome, 2013) starts from the important notion that flexible behavior is *context-dependent*. Context-dependent behavior involves selecting certain information in the environment, and mapping it to the appropriate behavior to achieve one's current ends. Even the exact same stimulus might require distinct action depending on the context. Mante et al. set out to study how cells in the prefrontal cortex could mediate this kind of flexibility. It is particularly interesting to think about these behaviors in a population of mixed-selectivity cells. If each cell's response properties are dependent upon multiple task parameters, how can those parameters be selectively employed to drive behavior?

To assess the properties of PFC cells in these contexts, Mante et al. trained monkeys to perform a complicated combination of tasks. The stimulus for the studies consisted of a field of dots, wherein either *motion* or *color* could be relevant for a given trial. It is well established that random motion of a set of dots produces no overall impression of movement direction. However, if degrees of "correlated" movement are introduced – say, 10% of the dots moving coherently to the left – pattern motion is both perceived and

reflected in direction-selective cells in the visual cortex (Britten et al., 1996). Moreover, percepts and neural responses are stronger the more correlated motion is included.

In addition to this well-understood paradigm, Mante et al. inserted another feature into the stimulus, namely the *proportion* of colors. So, the dots in the stimulus could vary in the percentage of them that were, e.g., red or green. There were thus two independent parameters of variation in the stimuli – the degree/direction of correlated motion, and the proportion of colors. The monkeys had to base their behavior on only one stimulus aspect per trial. Which was required in a given trial depended on a separate *context* cue. In "motion contexts," the monkeys were cued to make their choice based on the direction of predominant motion. In "color contexts" they were cued to decide based on the predominant color. Responses were saccades to locations. So, if motion was to the left in a motion trial, or the predominant color red in a color trial, the monkey would have to saccade left. Saccades to one direction were called "Choice 1" and in the other direction "Choice 2".

So, in total, there were four sources or "axes" of variation in the experiment. Two were in the stimulus, as described above. The other two were the context cue and the required behavior. The explanandum for the study was how the PFC organizes this information so as to produce the correct decision given the context.

There were two main parts to the study, first a PCA-based analysis of electrophysiological data, and second the construction of a model of PFC responses. The investigators measured cell responses to each trial in a widespread population of PFC cells. They then performed multiple regression on each cell's responses to see what task parameters drove the particular cell. Unsurprisingly, given the discussion so far, there was widespread mixed selectivity in the population. That is, the vast majority of cells had significant responses for multiple types of information. So, a given cell's responses would be mediated by some combination of the direction and degree of correlated motion, the preponderance of a certain color to a certain degree, the context, and the eventual choice. Selectivity for a certain task axis, then, is not the purview of some spatially or temporally distinct group of cells, but rather is distributed and mixed across the population.

The researchers performed PCA as a way of understanding population activity. As discussed above, PCA performed on the entire population subsequently allowed assessment of individual trials – at any given point, the activity of the population could be described via the weights of a subset of the principal components. In addition, however, the experimenters were interested in the combination of component weights that, specifically, correlated with the task axes. That is, which combination of principal component weightings particularly described the population's variation along each of the motion stimulus axis, the color axis, the context axis, and the choice axis? They used the

first 12 principal components, since these captured the majority of the variance in the population as a whole. They then described the task axes by looking for the combinations of weights on these components that best explained variance along each axis (Mante et al., 2013; extended data Fig 4.).[7]

Given this analysis, the experimenters then analyzed what happened across time in individual trials. The two panels in figure 2 below are both from the *motion context*, wherein the monkey had to make the choice based on the direction of motion. In this context, color information is *irrelevant* for the choice, but due to mixed selectivity across the population, color information still affects cell responses. The question is how the correct information is selected to drive behavior. The left panel depicts the choice axis (horizontal) and the motion axis (vertical) as described in the PC space. The center of the choice axis is taken to be the time when the stimulus is shown (i.e., before any decision process has begun). Each curve represents population responses to degrees and directions of correlated motion across several timesteps. As expected, when the monkey has to choose based on direction of motion, motion to the left drives the population to the "left" choice (Choice 1 here), whereas motion to the right drives the population to the "right" choice (Choice 2). The curves show how this process goes across each degree of correlation.

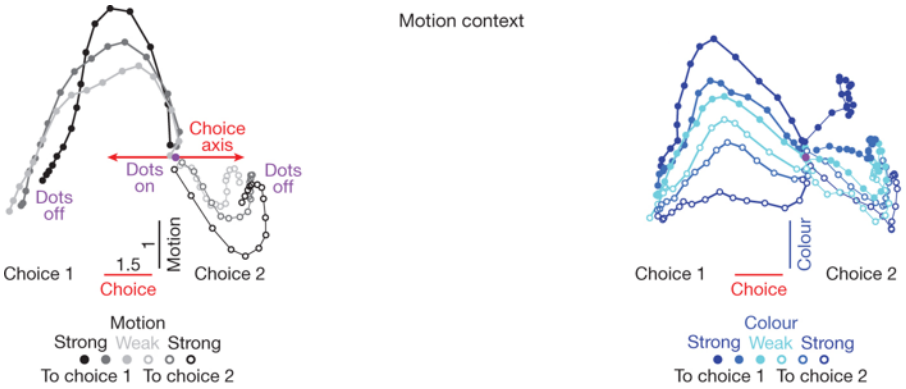

Figure 2. Population responses in the motion context. From Mante et al. (2013).

There is a vital difference between the left and right panels. Recall that this is the motion context, so only motion information is relevant. Unlike leftward motion, which only ever drives the population towards the "left" choice in this context, color information can result in *either* choice. That is, even strong predominance of one color can result in either Choice 1 or Choice 2. What this shows is that, while both color and motion affect population in each context, *only motion* moves the population along the choice axis in the motion context.

[7] The process of defining the "task axes" in the PC space is analytically complex. It involves creating a vector with normalized regression coefficients for each variable for each unit, combining them into a matrix, then constructing an orthogonal matrix via a technique known as "QR decomposition." See Mante et al. (2013, supplemental information; section 6.7) for full details.

Vitally, this pattern is *reversed* in the color context, as shown below. In this context, color predominance leads to a univocal choice, whereas motion can lead to either choice.

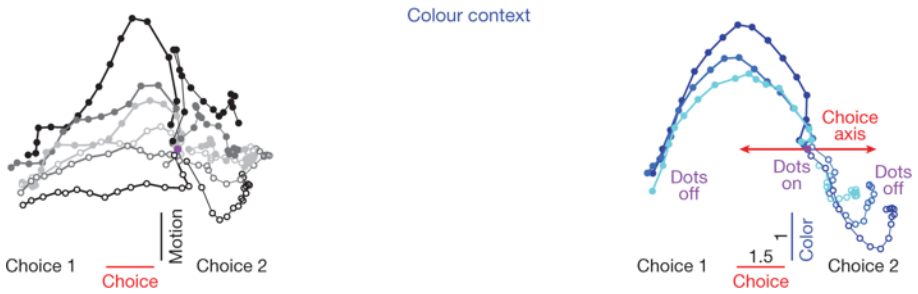

Figure 3. Population responses in the color context. From Mante et al. (2013).

So far, this is evidence of what the population is doing, not how it does it. Somehow, it selects which information will drive it along the choice axis. To attempt a mechanistic explanation for how this could come about, Mante et al. constructed and analyzed a recurrent neural network model. I lack space to go into the model in detail, but the important aspects for our purposes were (i) every cell received inputs corresponding to every task parameter, encouraging mixed selectivity, and (ii) the model was trained via feedback to produce a correct choice in the correct context, but was not given any information about *how* to produce that outcome.

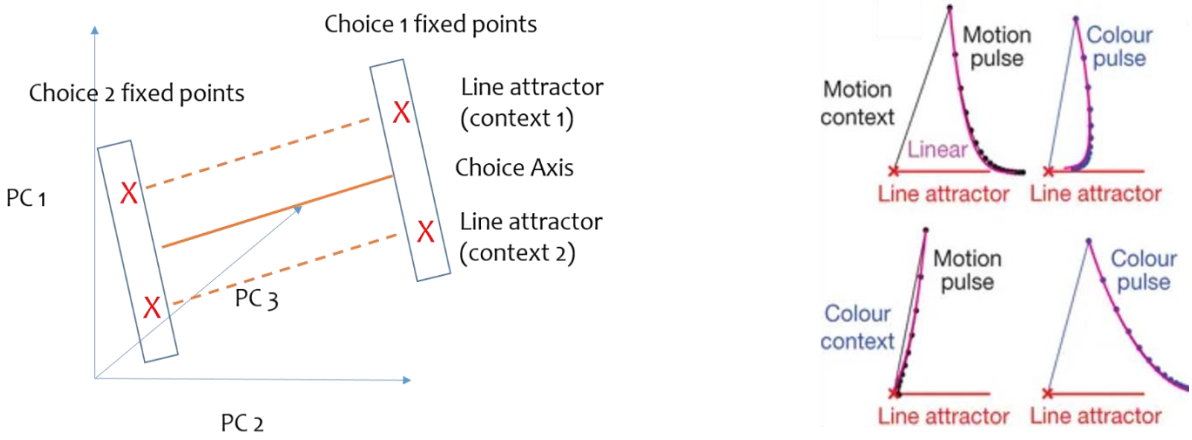

Figure 4. Left panel: an idealized rendering of the explanation based on line attractors, depicted in a simplified PC space. Right panel: different movements along line attractors in the motion and color contexts. From Mante et al. (2013).

Once the network was successfully trained, the experimenters could then manipulate the network in a fine-grained way to determine what happened as context changed. The full account of the system required distinguishing two "line attractors." These were trajectories

in the state space that were parallel to the axis of choice, but that differed between contexts (as visualized in simplified PC space in the left panel of figure 4). How they differed was shown by subjecting the model to brief pulses of motion and color. As it turns out, these stimuli were responded to differently in the motion versus color contexts. Think of the line attractor as a "path" that the system is on towards a conclusion. If the system was proceeding on the color context attractor, brief pulses of motion, when turned off, would result in the system moving back to where it had been on the attractor, whereas new color information would move it to a different place on the attractor. In the motion context, the pattern was reversed. This is shown in the right panel of figure 3. The key to note is that, while color may perturb the system in the motion context, it does not perturb the system in a direction along the line attractor. Motion information, however, does perturb the system towards a choice. Again, this pattern is reversed in the color context.

In sum, the model helps us understand how complicated, context-sensitive behaviors are parsed and implemented at the neural level. What the context cue is doing is setting up a certain dynamic organization in the system – in the color context, the system will proceed along one line attractor, whereas in the motion context it will proceed along the other. As Mante et al., note, despite the overlap of task-parameter selectivity in the population, "the network effectively implements two dynamical systems, one for each context" (2013; supplemental information, p. 17).

5. AH reconsidered

Shea argues that the Mante et al. analysis can be fit into his account. He suggests that there are two vehicles processing the inputs. One vehicle represents the context. Another vehicle has "two dimensions of variation," namely motion and color. The key function of the system is to take the inputs and translate them into a univocal action, picking the right information to guide the action. Shea takes the fact that information about the irrelevant dimension of the input "has little to no effect on which choice is programmed" (pp. 101-102) as evidence that the input is "transformed into a one-dimensional vector that drives behavior" (p. 184).

AH is committed to both injection mapping and causal isomorphism. As such, there has to be some fact of the matter about how to divide up the system into distinct, content-bearing parts, and to articulate the causal processes that transform contents between one vehicle and another. I contend, contra Shea, that there is no privileged way of dividing this system into distinct vehicles that causally interact in a stage-based manner. If that is the case, then AH fails.

The first aspect to focus on is mixed selectivity at the individual cell level. Recall that the task parameters are mixed at the individual cell level, meaning that cells have significant

responses across parameters. This means that one cannot ascribe contents corresponding to a given task dimensions to any single cell or group thereof. The second important aspect is how the analysis proceeded in light of these properties. Rather than attempting to decompose the population into spatially and temporally distinct vehicles, the researchers turned to analyzing the task parameters as differentiated by the way they are reflected in the activity of the whole population.

The dimensionality reduction is meant to portray how particular instantiations of activity across the system fit into the space of possible states of that system. And every individual instance of system behavior is described in terms of the principal component space, which is in turn constructed from the entire set of responses. So, given that (i) each cell is involved in all measures, and (ii) every individual instance is described in terms of a framework derived from the whole population, it is only at this level of analysis that the distinct task parameters can be differentiated.

Let's consider how this intersects with Shea's particular proposal. Recall that Shea posits one input vehicle with two dimensions of variation (motion and color), one context vehicle, and then a vehicle that transforms this content into a representation of the choice. I suggest that it is *arbitrary*, in both the spatial and temporal sense, to divide things up this way. Consider spatial decomposition first. Why should we say that there is one input vehicle with two dimensions of variation rather than two distinct vehicles? And, for that matter, why should we say that there are two vehicles, one for input and one for context, rather than one vehicle with *three* dimensions of variation? How would we tell?

In order for there to be a fact of the matter about spatial vehicle decomposition, one would have to find privileged parts within the system to which one content, rather than others, can be assigned. But this is precisely what the style of analysis in the Mante et al. paper disallows, due to its non-eliminable reliance on the entire population description for its explanatory framework.

Now consider the temporal aspect. Shea posits distinct stages of processing, on which the input contents are transformed into the output contents. Given the foregoing, there are a variety of problems with this. First, it implies that there are distinct vehicles underlying the inputs and the transformed contents. But we are in no better position with dividing up the vehicles temporally than we are spatially. What the temporal division requires is that there be a time at which one content is tokened, but the second content not tokened, and then a specific process that brings about that second content. But in the Mante et al. case, the task axes are constant features of the population response – every location in the PC space describes *some* place along the task axes. So, there is no temporal stage at which one content is tokened and the others aren't. Nor are there clear phases of transition between some represented contents and others. Given the global description of the system, the

representations of each of those elements is consistent and present throughout, and what changes is only where the system is in that representational structure.

Shea briefly considers the possibility of alternative divisions of the system, but thinks that we can rely on pragmatics to distinguish between them. Specifically with regards to the separation between inputs, he suggests that "Lumping the states of [context] and [input] together into a single vehicle on which behaviour is conditioned would be less explanatory… it would overlook an important aspect of how internal processing manages to perform the task" (p. 103). Suppose this is true at a purely pragmatic level – that the best way to understand the process is to think of it in terms of stages of information manipulation. The strongest claim this pragmatic move could defend is AC, which as we saw above is considerably weaker than AH.

A second hedge Shea occasionally makes use of is non-vicious content indeterminacy. He suggests that, in certain circumstances, we will be able to isolate a vehicle playing a certain role in an algorithm, but there will be limits on how determinately we can describe its contents. He thinks that, in this kind of situation, the causal and computational organization of the system will not privilege one ascription over another, and either will do a good job of explaining the behavior – hence, the indeterminacy is non-vicious. My worry, however, is not primarily with assignment of contents. It is whether the system can be spatially and temporally divided up in the way that AH requires. If vehicles are not arranged in such a manner as to support injective mapping and causal isomorphism, then AH fails, aside from questions about how to describe the contents.

Another potential response could appeal to a *level change*. The Mante et al. paper focuses on a single distributed process within the PFC, but Shea also discusses algorithms purportedly being played out across different brain areas, with transforms occurring in the causal relays between them. A potential response would be to admit that the system studied by Mante et al. is not decomposable into an algorithmic explanation, and propose instead to take the behavior of the *entire population* as a basic operation, that cannot be algorithmically decomposed any further, but that might be situated as one operation within a broader neural process/algorithm. Perhaps there are prior stages, outside of the population studied, that represent only the inputs and context, and what we see in the PFC population is a process by which these inputs are transformed into outputs, where this transformation doesn't need to be susceptible to a description in terms of AH.

A few things can be said against this strategy. First, pushing more and more of the processing into non-algorithmic dynamics undermines the *explanatory appeal* that AH is supposed to have. For Shea, what makes representational explanation, rather than mere causal explanation, relevant in cognitive science is that representational explanation gives us a perspicuous understanding of the organization of the system, which we cannot have

with non-representational explanation. But if much of the interesting work is being done by processes that do not work as AH suggests – and note that Mante et al. suggest their framework as a potentially general view of how the brain deals with context-sensitive behavior – then its general import is lessened.

Second, the level change strategy presupposes that, when we expand out to distinct brain areas, injection mapping and causal isomorphism will obtain. But this is not obviously the case. As noted, mixed selectivity and distribution of function is found widely throughout the brain, including in perceptual and subcortical areas. If isolating particular vehicles and contents in Mante et al.'s PFC population is difficult, there is no reason to suspect it to be easier as we expand out to other areas, at least not to the point where clear transformations over particular contents will be decipherable. Moreover, it isn't clear that positing such content transformation is *necessary*. As Hunt et al. (2017) suggest regarding the OFC circuit, it may be that different areas of the brain implement competitions within distinct reference frames – so the competition mediated by the OFC is in a value-frame, whereas dlPFC competitions might take place within a spatio-temporal reference frame. Put differently, it might be that if we go up a level we may get more of the same, now with brain networks representing complex combinations of information in a way that evolves dynamically towards a choice (Cisek & Thura, 2018; cf. Anderson, 2014; Burnston, 2019; Stanley, Gessel, & De Brigard, 2019). If so, then AH will not describe between-area processing either.

So, I conclude that, if these analyses are descriptive of population function, then AH fails for these systems. In the next section, I consider whether there is a kind of representational realism compatible with the falsity of AH. The answer will be a limited 'yes'.

6. Representational Realism?

I have argued that, at least if the kinds of analyses I have discussed reveal the functional properties of the PFC, then AH is false for that system. There is a very flexible sense in which responses in the system may correlate with contents posited in an algorithm, but this at best supports AC rather than AH. In this section, I consider the upshot of this result for thinking about representation.

Recall that Shea does not distinguish vehicle realism from AH, treating them as something of a package deal. I agree that vehicle realism is non-negotiable for realism about representation, but I suggest that one can endorse it without assenting to AH. I propose that there *are* distinctive vehicles for distinct contents, but these are not organized into sequential steps, and indeed overlap considerably at the level of individual parts.

To a certain extent, this requires ambiguating on "distinctive." AH requires that *distinct physical parts* correspond to distinct contents, and interact at distinct causal/temporal stages.

I suggest that this is simply not a requirement for vehicle realism. The kinds of trajectories and competitive processes discussed in the two case studies can occur even if the relevant aspects of content are completely mixed at the individual cell level, and if *all* of the cells in the population contribute to its function. What matters for the explanations given above, especially as articulated in the Mante et al. analysis, is that the population has distinctive patterns of activity that capture the relevant variance in the task, in a way that maps the right inputs to the right outputs. So, my proposal:

- Vehicle realism: A neural system has distinct vehicles corresponding to contents a…n if (i) variation in a...n have distinct effects on the dynamics of the system, and (ii) those distinct effects result in the fulfillment of a task function.

What "distinct" means on this rendering is that the system is set up in such a way that different aspects of the task (including both behavioral and environmental variables) have independent effects in the system. This is shown in several ways in the Mante et al. study. Consider figure 2 again. What it shows is that motion and color information influence the system differently, and that no matter the choice context, variation in motion and color are tracked within the system. The context, however, changes the which of those effects will in turn drive the system along the axis of choice – that is, context signals push the system onto one line attractor versus the other, regardless of the stimulus value. Choice 1 versus Choice 2 outcomes are reflected in the fixed points at which the process can end. So, motion information, color information, and information about task context all have distinct effects within the system, and these combined effects result in the system approaching a fixed point that constitutes the choice.

This version of vehicle realism is compatible with *complete overlap* between the spatial parts that represent the distinct contents (cf. Levy & Bechtel, 2016). But this in no way impugns the independent effects of the distinct parameters. According to the explanation, and as exhibited in the neural network model, it is a particular learning history, along with synaptic modification of the connections in the population, that sets up the state space. That is, a particular physical arrangement of the system is what underlies its dynamics, and hence what allows for the distinct effects of variation in the task parameters. Moreover, this kind of posit is not explanatorily neutral. If Mante et al. are right, then the fact that each task parameter has distinct effects on the system, and that the system is driven in the way that it is by those parameters, is *why* the organism can choose the right stimulus aspect to shape their action in the right context.

My version of vehicle realism agrees with Shea that reference to task function is non-eliminable in positing representation. As Shea notes, reference to task function is necessary to specify what kind of properties are important for the organism. The analysis in this section further supports this view. Consider just how expansive the state space of a system

like the PFC is. In principle, it has a dimension corresponding to the level of activation of every single neuron. What the researchers presume is that the functionally relevant aspects of the system are those that allow it to separate and respond to the appropriate task axes. This kind of analysis is not possible without an understanding of the environmental and task structure in which the organism is operating.

So, much of Shea's analysis can be salvaged without having to endorse AH. My version of vehicle realism, combined with AC, is proposed as an alternative. What remains is to ask whether this is a robust enough notion of representation to be worthy of the name. Much of the recent discussion surrounding representation in cognitive science has been focused on the "job-description" challenge – in virtue of what are representational posits truly explanatory (Morgan, 2013; Ramsey, 2007)? If one does not believe that posits of representations meet this challenge, a range of positions from eliminativism (Hutto & Myin, 2014) to pragmatism (Coehlo Mollo, 2017; Egan, 2014b) to fictionalism (Sprevak, 2013) are available. A lot of ground has been covered here, which I don't intend on re-treading. I take three of the most interesting forms of responses to the challenge to be those that posit a certain kind of *structure* for representations (Isaac, 2013), those that tie the function of representation to serving as an internal model (Grush, 2004), and those that, as discussed in section 2, appeal to the architecture of the system.

These views are, in my opinion, largely on the right track. If I am right in the foregoing, however, then the reason these views are right cannot be because AH obtains. What I suggest is that the population studied by Mante et al., as a whole, exhibits a structural mapping to the important sources of variance in the task – both its environmental and behavioral variables.[8] As described in the PC space, the population has distinctive activity patterns corresponding to variation along the task axes. So, changes in the relevant task variables are systematically reflected in changes to the system's behavior. I thus suggest that, since (i) the response along task axes is due to the physical organization of the system, and (ii) that organization is what enables the system to produce the appropriate task function, the dimensionality-reduction analyses are sufficient to ground the description of the system in representational terms.[9]

---

[8] I am stressing the structural notion of representation here, but the correlational notion is important for the Mante et al. study as well: the significant regression coefficients for selectivity to multiple task parameters in individual cells is a foundational part of the analysis. Selectivity is mixed throughout the population, so individual cells cannot be attributed univocal contents. But correlations do not have to be univocal to ground the informational notion of representation.

[9] This view, which posits representational realism despite divorcing it from a particular algorithmic approach, I believe, fits well with Nanay's (2018) recently proposed "entity realism" (cf. Plebe & De La Cruz, 2018; Thomson & Piccinini, 2018). There is an interesting question about what this does to specifically *computational* explanation. As noted, Shea conjoins these two while I dissociate them. A range of possibilities exist here. It may, for instance, be the case that neural representations of the type offered here are coherent with multiple algorithmic descriptions, without there being a fact of the matter

Finally, while I think my version of vehicle realism supports realism about certain forms of representation, this is a far cry from legitimizing *all* representation talk. The kind of vehicles I'm discussing are tightly coupled with, if not completely determined by, patterns of variation in the environment. It is possible that my view of representation will support only a limited realism, where what the brain represents is heavily conditioned on interaction with the environment, and is action-oriented (Anderson, 2014). One can believe this and still be a pragmatist or fictionalist about the kinds of determinate, categorical, linguistically expressible contents often discussed in the philosophical literature, about propositional attitudes, etc. (My own view on the matter is slightly more complex; see Burnston, 2017a, 2017b.) I take no official stand here.

7. Conclusion

I have attempted to isolate an idea, algorithmic homuncularism, which is widely prevalent in thinking about cognitive neuroscience, and clearly articulated in Shea's recent book. While AH has indeed been influential in cognitive neuroscience, and has significant philosophical appeal, certain kinds of complexity in the functional profiles of neural systems like the PFC – as well as in the analysis strategies scientists use to explore them – militate against it. I have argued that, even if AH fails, this is no argument for eliminativism about representation.

REFERENCES

Akam, T., & Kullmann, D. M. (2014). Oscillatory multiplexing of population codes for selective communication in the mammalian brain. *Nature Reviews Neuroscience, 15*(2), 111-122.

Anderson, M. L. (2014). *After phrenology: Neural reuse and the interactive brain*. Cambridge, MA: MIT Press.

Bechtel, W. (1998). Representations and cognitive explanations: Assessing the dynamicist's challenge in cognitive science. *Cognitive science, 22*(3), 295-318.

Bechtel, W. (2014). Investigating neural representations: the tale of place cells. *Synthese, 193*(5), 1287-1321. doi:10.1007/s11229-014-0480-8

Boone, W., & Piccinini, G. (2016). The cognitive neuroscience revolution. *Synthese, 193*(5), 1509-1534.

about which one is right. And what computational description of a vehicle's operation is best may vary with context (Burnston, 2016b). Thus, it is possible to be an entity realist about representations while having a different view on computational explanation. One particularly appealing position, in light of the above, is Chirimuuta's (forthcoming) recently developed "perspectival" realism about computational posits, but arguing for this is beyond the scope of this paper.

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., & Movshon, J. A. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis Neurosci, 13*, 87-100.

Burnston, D. C. (2016a). Computational neuroscience and localized neural function. *Synthese, 193*(12), 3741-3762.

Burnston, D. C. (2016b). A contextualist approach to functional localization in the brain. *Biology & Philosophy, 31*(4), 527-550.

Burnston, D. C. (2017a). Cognitive penetration and the cognition–perception interface. *Synthese, 194*(9), 3645-3668. doi:10.1007/s11229-016-1116-y

Burnston, D. C. (2017b). Interface problems in the explanation of action. *Philosophical Explorations, 20*(2), 242-258.

Burnston, D. C. (2017c). Real Patterns in Biological Explanation. *Philosophy of Science, 84*(5), 879-891. doi:10.1086/693964

Burnston, D. C. (forthcoming). Getting over Atomism: Functional Decomposition in Complex Neural Systems. *The British Journal for the Philosophy of Science*.

Canolty, R. T., & Knight, R. T. (2010). The functional role of cross-frequency coupling. *Trends in Cognitive Sciences, 14*(11), 506-515.

Cao, R. (2011). A teleosemantic approach to information in the brain. *Biology & Philosophy, 27*(1), 49-71.

Chemero, A., & Silberstein, M. (2008). After the Philosophy of Mind: Replacing Scholasticism with Science. *Philosophy of Science, 75*(1), 1-27.

Chirimuuta, M. (2014). Minimal models and canonical neural computations: the distinctness of computational explanation in neuroscience. *Synthese, 191*(2), 127-153.

Chirimuuta, M. (2017). Explanation in computational neuroscience: Causal and non-causal. *The British Journal for the Philosophy of Science*.

Chirimuuta, M. Forthcoming. 'Charting the Heraclitean Brain: Perspectivism and Simplification in Models of the Motor Cortex.' in Michela Massimi and Casey McCoy (eds.), *Understanding Perspectivism: Scientific Challenges and Methodological Prospects* (Routledge: New York).

Cisek, P., & Thura, D. (2018). Neural Circuits for Action Selection. *Reach-to-Grasp Behavior: Brain, Behavior, and Modelling Across the Life Span*.

Clark, A. (1998). *Being there: Putting brain, body, and world together again*: MIT press.

Coelho Mollo, D. (2017). Content Pragmatism Defended. *Topoi, 39*(1), 103-113.

Cummins, R. (1985). The nature of psychological explanation.

Cummins, R. (2000). How does it work?" versus" what are the laws?": Two conceptions of psychological explanation. *Explanation and cognition*, 117-144.

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

Egan, F. (2014a). Function-Theoretic Explanation and the Search for Neural Mechanisms. In D. Kaplan (Ed.), *Integrating Mind and Brain Science: Mechanistic Perspectives and Beyond*.

Egan, F. (2014b). How to think about mental content. *Philosophical Studies, 170*(1), 115-135.
Elber-Dorozko, L., & Shagrir, O. (2019). Integrating computation into the mechanistic hierarchy in the cognitive and neural sciences. *Synthese*.
Fodor, J. A. (1975). *The language of thought* (Vol. 5): Harvard University Press.
Grush, R. (2004). The emulation theory of representation : Motor control , imagery , and perception. 377-442.
Hunt, L. T., Behrens, T. E., Hosokawa, T., Wallis, J. D., & Kennerley, S. W. (2015). Capturing the temporal evolution of choice across prefrontal cortex. *eLife, 4*.
Hunt, L. T., & Hayden, B. Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. *Nat Rev Neurosci, 18*(3), 172-182. doi:10.1038/nrn.2017.7
Hutto, D. D., & Myin, E. (2014). Neural representations not needed-no more pleas, please. *Phenomenology and the Cognitive Sciences, 13*(2), 241-256.
Kästner, L., & Haueis, P. (2019). Discovering Patterns: On the Norms of Mechanistic Inquiry. *Erkenntnis*, 1-26.
Isaac, A. M. (2013). Objective similarity and mental representation. *Australasian Journal of Philosophy, 91*(4), 683-704.
Levy, A., & Bechtel, W. (2016). Towards Mechanism 2.0: Expanding the Scope of Mechanistic Explanation.
Maley, C. J. (2018). Toward analog neural computation. *Minds and Machines, 28*(1), 77-91.
Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature, 503*(7474), 78-84.
Morgan, A. (2013). Representations gone mental. *Synthese, 191*(2), 213-244.
Millhouse, T. (forthcoming). "Compressibility and the Reality of Patterns." *Philosophy of Science*.
Millikan, R. G. (1989). Biosemantics. *The Journal of Philosophy, 86*(6), 281-297.
Nanay, B. (2019). Entity Realism About Mental Representations. *Erkenntnis*, 1-17.
Padoa-Schioppa, C., & Conen, K. E. (2017). Orbitofrontal Cortex: A Neural Circuit for Economic Decisions. *Neuron, 96*(4), 736-754.
Panzeri, S., Macke, J. H., Gross, J., & Kayser, C. (2015). Neural population coding: combining insights from microscopic and mass signals. *Trends in Cognitive Sciences, 19*(3), 162-172.
Piccinini, G. (2007). Computing mechanisms. *Philosophy of Science, 74*(4), 501-526.
Piccinini, G. (2008). Computation without representation. *Philosophical Studies, 137*(2), 205-241.
Piccinini, G., & Bahar, S. (2012). Neural computation and the computational theory of cognition. *Cognitive science*.
Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: functional analyses as mechanism sketches. *Synthese, 183*(3), 283-311. doi:10.1007/s11229-011-9898-4
Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of Biological Physics, 37*(1), 1-38.

Piccinini, G., & Shagrir, O. (2014). Foundations of computational neuroscience. *Current opinion in neurobiology, 25*, 25-30.
Plebe, A., & De La Cruz, V. M. (2018). Neural Representations Beyond "Plus X". *Minds and Machines, 28*(1), 93-117.
Povich, M. (2015). Mechanisms and model-based functional magnetic resonance imaging. *Philosophy of Science, 82*(5), 1035-1046.
Povich, M. (2019). Model-based cognitive neuroscience: Multifield mechanistic integration in practice. *Theory & Psychology, 29*(5), 640-656.
Ramsey, W. M. (2007). *Representation reconsidered*: Cambridge University Press.
Rathkopf, C. (2017). What Kind of Information is Brain Information? *Topoi, 39*(1), 95-102.
Rescorla, M. (2014). The causal relevance of content to computation. *Philosophy and Phenomenological Research, 88*(1), 173-208.
Rich, E. L., & Wallis, J. D. (2016). Decoding subjective decisions from orbitofrontal cortex. *Nature neuroscience, 19*(7), 973.
Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature, 497*(7451), 585-590.
Rowlands, M. (2009). Situated representation. In M. Aydede & P. Robbins (Eds.), *The Cambridge handbook of situated cognition* (pp. 117-133). Cambridge: Cambridge University Press.
Rupert, R. D. (2018). Representation and mental representation. *Philosophical Explorations, 21*(2), 204-225.
Saez, A., Rigotti, M., Ostojic, S., Fusi, S., & Salzman, C. (2015). Abstract context representations in primate amygdala and prefrontal cortex. *Neuron, 87*(4), 869-881.
Shagrir, O. (2006). Why we view the brain as a computer. 393-416.
Shagrir, O. (2010). Brains as analog-model computers. *Studies in History and Philosophy of Science Part A, 41*(3), 271-279.
Shagrir, O. (2012a). Computation, Implementation, Cognition. *Minds and Machines, 22*(2), 137-148.
Shagrir, O. (2012b). Structural Representations and the Brain. *The British Journal for the Philosophy of Science, 63*(3), 519-545.
Sprevak, M. (2013). Fictionalism about neural representations. *The Monist, 96*(4), 539-560.
Stanley, M. L., Gessell, B., & De Brigard, F. (2019). Network Modularity as a Foundation for Neural Reuse. *Philosophy of Science, 86*(1), 23-46.
Thomson, E., & Piccinini, G. (2018). Neural representations observed. *Minds and Machines, 28*(1), 191-235.
Watrous, A. J., Fell, J., Ekstrom, A. D., & Axmacher, N. (2015). More than spikes: common oscillatory mechanisms for content specific neural representations during perception and memory. *Current opinion in neurobiology, 31*, 33-39.
Weiskopf, D. A. (2015). The explanatory autonomy of cognitive models. *Integrating Psychology and Neuroscience: Prospects and Problems. Oxford University Press, Oxford*.