Neural Representations Unobserved - or: a dilemma for the cognitive neuroscience revolution

Marco Facchin, Ph.D. [**corresponding author**]
Research Fellow
Istituto Universitario di Studi Superiori IUSS Pavia
Linguistics & Philosophy IUSS Center, Department of Human and Life Sciences; Pavia, Italy
Palazzo del Broletto, Piazza della Vittoria n. 15, 27100, Pavia
marco.facchin@iusspavia.it
https://orcid.org/0000-0001-5753-987

Abstract:

Neural structural representations are cerebral map- or model-like structures that structurally resemble what they represent. These representations are absolutely central to the "cognitive neuroscience revolution", as they are the only type of representation compatible with the revolutionaries' mechanistic commitments. Crucially, however, these very same commitments entail that structural representations can be observed in the swirl of neuronal activity. Here, I argue that no structural representations have been observed being present in our neuronal activity, no matter the spatiotemporal scale of observation. My argument begins by introducing the "cognitive neuroscience revolution" (**§1**) and sketching a prominent, widely adopted account of structural representations (**§2**). Then, I will consult various reports that describe our neuronal activity at various spatiotemporal scales, arguing that none of them reports the presence of structural representations (**§3**). After having deflected certain intuitive objections to my analysis (**§4**), I will conclude that, in the absence of neural structural representations, representationalism and mechanism can't go together, and so the "cognitive neuroscience revolution" is forced to abandon one of its commitments (**§5**).

Keywords: Structural representations, Eliminativism, Neuroscience, Mechanistic explanation, Neurocognitive revolution.

Declarations:

Founding: Not applicable

Conflict of Interests/competing interests: The author declares no conflict of interests.

Availability of data and Material: Not applicable.

Code availability: Not applicable.

Authors' contribution: Marco Facchin is the sole author of the paper.

Acknowledgements: Thanks to (in random order) Marco Viola, Davide Coraci, Arianna Beghetto, Jonny Lee and Sanja Sreckovic for having read and commented upon several previous

poorly written and half-baked versions of this paper. Thanks to (again, in random order) Erik Thomson, Bryce Huebner and Carl Sachs for an extremely insightful exchange via Twitter on cortical maps and structural representations. Thanks also to the public of the 4th International Conference in Philosophy of Mind held in Braga for their stimulating questions.

Neural Representations Unobserved - or: a dilemma for the cognitive neuroscience revolution

Abstract:

Neural structural representations are cerebral map- or model-like structures that structurally resemble what they represent. These representations are absolutely central to the "cognitive neuroscience revolution", as they are the only type of representation compatible with the revolutionaries' mechanistic commitments. Crucially, however, these very same commitments entail that structural representations can be observed in the swirl of neuronal activity. Here, I argue that no structural representations have been observed being present in our neuronal activity, no matter the spatiotemporal scale of observation. My argument begins by introducing the "cognitive neuroscience revolution" (**§1**) and sketching a prominent, widely adopted account of structural representations (**§2**). Then, I will consult various reports that describe our neuronal activity at various spatiotemporal scales, arguing that none of them reports the presence of structural representations (**§3**). After having deflected certain intuitive objections to my analysis (**§4**), I will conclude that, in the absence of neural structural representations, representationalism and mechanism can't go together, and so the "cognitive neuroscience revolution" is forced to abandon one of its commitments (**§5**).

Keywords: Structural representations, Eliminativism, Neuroscience, Mechanistic explanation, Neurocognitive revolution.

1 - Introduction: neural structural representations and the cognitive neuroscience revolution

Representations remain as central to cognitive science as elusive to our understanding (Villaroja 2017; Favela & Machery forthcomin*g*). Philosophers invested in the "cognitive neuroscience revolution" (Boone and Piccinini 2016), however, argue that cognitive *neuro*science[1] operates upon a stable concept of *neural* representation. In their view, cognitive neuroscience provides us with a concept of neural representations as map- or model- like structures that represent their targets by resembling them in a particular, *structural* way. Call these representations neural structural representations - NSRs for short (see Gładziejewski 2015; 2016; Gładziejewski & Miłkowski 2017; Williams 2017; Williams and Colling 2017; Wiese 2016, 2017; Morgan & Piccinini 2018; Piccinini 2020a, 2020b, 2022).

Prima Facie, contemporary cognitive neuroscience relies heavily on NSRs. The spatial navigational skills of rats are explained by appealing to a cognitive *map* hosted in the rat's hippocampus (cf. O'Keefe & Nadel 1978; Moser *et al*. 2018). Motor control is accounted for in terms of various *models* computing and controlling the relevant motor trajectories (Pickering & Clark 2014; Mcnamee & Wolpert 2019). Moreover, the very same *models* might underpin social cognition (Haruno *et al*. 2003). The "mirror" property of many neurons is increasingly interpreted in terms

[1] Here,"cognitive neuroscience" and "cognitive science" will refer only to mainstream approaches - that is, representational and computational - in the respective disciplines. For non-mainstream alternatives, see (Kelso 1995; Chemero 2009; Anderson 2014; Bruineberg & Rietveld 2019; Van der Weel *et al*. 2022).

of inner *models* allowing to simulate actions (Kilner *et al*. 2007; Csibra 2008) and emotions (Rizzolatti & Sinigaglia 2023) offline. Popular neurocomputational frameworks such as predictive processing cast all brain functions as operations on complex, multifaceted statistical *models* of the environment (cf. Buckley *et al*. 2017).[2] More generally, the idea that inner *models* are the only way in which an agent can make sense and control the flux of input the environment bombards the agent with is gaining momentum (Seth 2015; Brette 2019). The cognitive centrality of inner *models* is further confirmed by a host of neurorobotic experiments (Tani 2007; 2016) and neurocomputational models (cf. Ha & Schmidhuber 2018a, 2018b, Poldrack 2020). And so, whilst such map- and model- like structures are in no way the *only* type of representational structure cognitive neuroscientist invoke (cf. Barack & Krakauer 2021; Backer *et al*. 2022; Frisby *et al*. 2023), it is undeniable that they do play a large explanatory role in contemporary cognitive science. For proponents of the "cognitive neuroscience revolution", however, NSRs are not "just" important - they are *central* to the success of cognitive neuroscience.

This is because supporters of the "cognitive neuroscience revolution" claim that cognitive neuroscience is deeply committed to a *mechanistic* explanatory strategy (see Gładziejewski 2015; Boone & Piccinini 2016; Williams & Colling 2017; Piccinini 2020a).[3] In such a view, to explain cognitive capacities (and their behavioral manifestations) is to identify and describe the physical mechanism responsible for such capacities (and behavioral manifestations). Otherwise put: to explain a cognitive capacity (or a behavioral manifestation thereof) is to identify and describe a set of organized physical components whose causal interaction constitutes the cognitive capacity in question (or causes the relevant behavioral manifestation; see Craver 2007; Bechtel 2008). Crucially, mechanistic explanations are (at least partially) *ontic* explanations. Their explanantia are not (only) statements concerning mechanisms, but the *actual* mechanisms "in flesh and blood", so to speak (cf. Craver 2007, p. 27; Illari 2013).

Now, if the representational, content-based explanations of cognitive (neuro)science are mechanistic explanations, then it seems that (neural) representations must be real and literal components of our (neuro)cognitive mechanisms, whose content must literally and really be causally efficacious within the mechanisms's inner functional economy. And this entailment is *prima facie* highly problematic. For, it is quite natural to think that representational contents are causally inert. All the heavy causal lifting seems done by the representational vehicles - the physical structures "doing" the representing by "carrying" the contents around - rather than the contents themselves (e.g. Egan 2020). So, doesn't the mechanistic approach to explanation prevent us from restoring to content-based, representational explanations?

No, it doesn't - or so the proponents of the cognitive neuroscience revolution claim. For, structural representations are underpinned by representational vehicles whose physical shape is not just casually potent, but also *semantically relevant*. This is because the physical shape of the vehicles, and the particular way in which they resemble their targets, determines what these vehicles represent. In the case of NSRs, then, semantic content and vehicular shape are at least largely overlapping, if not the exact same thing (Williams & Colling 2017; Lee 2019; Piccinini 2022). In

[2] Even if Predictive Processing also admits non-representational, "Model free" interpretations (Downey 2018; Facchin 2021a). These interpretations, however, remain fairly unorthodox.

[3] But see (Silberstein & Chemero 2013; Silberstein 2021) for a diverging opinion.

this way, semantic contents are able to play an active causal role within our neurocognitive mechanisms, and are thus able to play a genuine explanatory role in mechanistic explanations (cf. O'Brien 2015).

But notice how, given that NSRs are bona fide components of neurocognitive mechanisms, they must be observable and manipulable as any other component of said mechanisms. Proponents of the neurocognitive revolution agree - either implicitly (see Williams 2017) or explicitly (Piccinini 2020) - that this is the case. This means that, at least when it comes to NSRs, we can circumvent the seemingly never-ending debate concerning the reality of internal representations (cf. Ramsey 2007; Hutto & Myin 2013; Segundo-Ortin & Hutto 2021; Anderson & Champion 2022). To determine whether NSRs are real, one just needs to take a peek inside the neurocognitive system and see whether NSRs - or, more accurately, NSRs-supporting vehicles[4] - can be found (cf. Bechtel 2008, 2014; Thomson & Piccinini 2018; Piccinini 2020a; Facchin 2021a). For simplicity, let me refer to NSRs supporting vehicles as NSRVs.

The aim of this paper is to take one such peek. As its title might have revealed, I will argue that no NSRVs can be observed. My analysis will unfold as follows. (**§2**) introduces a widely accepted account of structural representations, focusing on the constraint it places on representational vehicles. (**§3**) considers whether neuronal vehicles satisfy these constraints. (**§3.1**) focuses on individual neuronal responses. (**§3.2**) focuses on neuronal maps. (**§3.3**) focuses on activation spaces In all these cases, I will argue that the relevant candidate vehicles cannot satisfy the constraints introduced in (**§2**). Hence, these vehicles cannot be NSRVs. (**§4**) anticipates some objections. (**§5**) considers the implications of my verdict for cognitive neuroscience, concluding the paper.

2 - A standard account of (neural) structural representations

Informally described, structural representations are model- or map- like structures which represent their targets (i.e. what the representation is "aimed at") by being structurally similar to them. Cartographic maps are paradigmatic examples of structural representations since they represent a terrain by replicating the terrain spatial structure with their own spatial structure: if location *a* is west of location *b*, then the map will display the point standing for *a* left of the point standing for *b*. Now, how can this intuitive, but imprecise, idea of a structural representation be made more rigorous?

Paweł Gładziejewski (2015; 2016) offers a nowadays standardly accepted philosophical analysis of structural representations:

> Within a system S, a vehicle V is the vehicle of a structural representation of a target T if and only if:
> **(1) Structural similarity:** V is structurally similar to T; &
> **(2) Action Guidance:** The structural similarity in **(1)** allows V to guide S's action in regards to T; &

[4] This caveat is actually important: NSRs proper are *relations* between neural vehicles and their targets, so they can't be observed *just* by observing neural goings on. At best, then, observing neural goings lets us see one *relatum*, that is, the relevant representational vehicles (the NSRV).

> **(3) Decouplability:** **(2)** can obtain even when V is decoupled from T; &
> **(4) Error Detection:** S can detect the representational errors V generates

There is much to say about **(1)**-**(4)**, both as individually and as a whole. One first important thing to notice is that they all concern structural representations *in general* - they're not specific to NSRs. This is a good thing, as it allows me to explain **(1)**-**(4)** in terms of structural representations everyone is familiar with, such as maps. The step from structural representation in general to NSRs can then be easily made by placing an appropriate restriction on the physical medium realizing the vehicles: vehicles must be realized by neurons - or, more precisely, by patterns of neuronal activities.

Secondly, **(1)**-**(4)** all concern the *vehicle* of a structural representation. Consider, for example, the physical support underpinning a cartographic map. It is *that support* - that is, the representational vehicle - that **(1)** is structurally similar to the mapped terrain, **(2)** is used to guide our actions (e.g. in traversing said terrain), **(3)** can guide our actions when we're decoupled from that terrain (e.g. allowing us to plan the way ahead), and **(4)** whose usage allows us to detect its eventual representational errors (e.g. by noticing that it leads us systematically astray). So, **(1)**-**(4)** specify the relevant vehicular features underpinning structural representations. Notice also that, since **(1)**-**(4)** are imposed in conjunction, the vehicles underpinning structural representations must satisfy all of them. I will now examine each condition in turn, focusing in particular on **(1)** and **(2)**, as they will be extremely important throughout the entirety of (**§3**).

Condition **(1)** requires the representational vehicle V to be structurally similar to the represented target T. The relevant structural similarity relation can be unpacked in a number of ways. Like Gładziejewski, I chose a very liberal unpacking.[5] Choosing such a liberal unpacking makes **(1)** *easier* to satisfy - and so, NSRVs easier to spot. Thus, this is the relevant charitable interpretation of NSRs in the present context. In this view:

> V is structurally similar to T if and only if:
> **(a)** There is a one-to-one mapping from some vehicle constituents ($v_a...v_n$) of V to some target constituents ($t_a...t_n$) of T; &
> **(b)** There is one relation R holding among the vehicle constituents of V and one relation R^* holding among the target constituents of T such that, for all the vehicle constituents satisfying **(a)**: $(v_a R v_b) \rightarrow (t_a R^* t_b)$. (cf. O'Brien and Opie 2004).

(a) imposes a *one-to-one* mapping from some relevant physical bits and pieces of the vehicle V (i.e. vehicle constituents) to some bits and pieces of the target T (i.e. the target constituents). I won't pose any restriction on what may count as a vehicle constituent - everything may be vehicle constituent, provided that it is a material constituent of a vehicle. For the sake of simplicity, however, I won't consider here arbitrary, gendermarried or otherwise "unnatural" way of carving up vehicles: whilst "unnatural" mappings alway allow to find a structural similarity (cf. McLendon 1955), it is very doubtful our neurocognitive systems *care* about them - they won't be, as Shea (2018) usefully puts it, *exploitable* by our neurocognitive system. Also, again for the sake of

[5] For less liberal views, see (Swoyer 1991; Isaac 2013)

simplicity, I'll always assume that the mapping in **(a)** is "subscript preserving": v_a maps onto t_a, v_b maps onto t_b, ... and v_n maps onto t_n.

(b) forces V and T to share the same inner relational structure: if a relevant relation *R* holds between v_a and v_b, then a relevant relation R^* holds between t_a and t_b. Notice that **(b)** mentions *one* relation in V and *one* in T. So, in order for **(b)** to obtain the relations preserved by the mapping in **(a)** needs to be constant on both sides of the mapping. By this I *do not* mean that *R* and R^* must be the *same* relation.[6] I mean something different - namely that the relation at one side of the similarity cannot "switch". Thus, if $(v_aRv_b) \rightarrow (t_aR^*t_b)$ but $(v_cRv_d) \rightarrow (t_cR^{**}t_d)$, then **(b)** fails to obtain. Imagine a map representing the distance between some cities in a region in terms of distances between them, *and also* the distances between other cities in the same region only in terms of the *colors* used to represent the cities (e.g. cities represented in *darker* colors are further apart than cities represented in *lighter* colors). Such a map *would not* count as a structural representation according to Gładziejewski's analysis (and it would also be *really* hard to use).

Crucially, conditions **(a)** and **(b)** determine the relevant semantic properties of structural representations. They determine what a vehicle V represents.[7] In structural representations, v_a *represents* t_a, and the fact that v_aRv_b *represents that* $t_aR^*t_b$ (e.g. Shea 2018). Thus **(a)** and **(b)** - that is, **(1)** - are the reasons why the physical shape of the representational vehicles of structural representations are imbued with their semantic properties (Williams and Colling 2017).

Notice how **(1)** entails that structural representations have a specific form of semantic transparency. Since the mapping in **(a)** is one-to-one and **(b)** operates only on *one* relation for V and one relation for T, then it is always possible to interpret all the "v_xRv_y" univocally and transparently: v_aRv_b can only represent $t_aR^*t_b$. Notice that since structural representations are transparent, their content is neither disjunctive nor indeterminate: v_aRv_b *represents that* - and only that - $t_aR^*t_b$. Were it to represent something disjunctive or indeterminate - say, something like ($t_aR^*t_b$ *or* $t_{a^*}R^*t_b$) or ($t_aR^*t_b$ *or* $t_aR^{**}t_b$) - then either **(a)** or **(b)** (and so, **(1)**) would fail to obtain.

Notice further that the fact that the obtaining of **(1)** determines the semantic properties of structural representations does not entail that **(1)** is the *content-grounding relation* in virtue of which V is a representation of T (cf. Von Eckart 1996). The structural similarity in **(1)** *need not* be what "makes" a vehicle contentful; it need not be the factor in virtue of which a vehicle comes to represent a target and thus have a content - even if, according to some, it may (see Cummins 1996; Lee 2018, 2021). V may come to represent T for different reasons - for example, in virtue of its informational linkage with T or its proper functions in regard to T (Ramsey 2016; Neander 2017; Wiese 2017; Piccinini 2020a, b, 2022). The structural similarity in **(1)** *necessarily* determines only *how* T is represented. That is, if v_aRv_b is part of that similarity, then T is represented as being such that $t_aR^*t_b$ is the case. Compare: whilst the way in which a map represents a territory is set by the way the two are structurally similar - that it, the relevant structural similarity in **(1)** - presumably the map does not represent the territory *in virtue of* that structural similarity, but rather in virtue of

[6] Even if it is possible that $R=R^*$ - after all, when cartographic maps are involved, spatial relations are preserved on both sides of the mapping.

[7] Or, minimally, *some* of the relevant semantic properties of structural representations - other ingredients may be necessary to account for all the semantic properties of structural representations. To give but one example Shea (2018) argues a teleological component is needed to.

certain map-involving social practices. At any rate here I will stay neutral on whether **(1)** is the content grounding relation of NSRs, and determining such an issue has no bearing on my arguments.

Condition **(2)** is satisfied when the structural similarity in **(1)** guides the actions of a system S that are "aimed at" T. When this happens, S's odds of success are sensitive to the *quality* of the similarity holding between V and T (see Shea 2018, p.142). The more V structurally resembles T, the higher S's odds of non-accidental success; and, the lower the quality of the resemblance, the lower S's odds. *Ceteris paribus*[8], the better the map resembles the terrain, the more one is able to traverse it. The worse their resemblance, the more one is likely to get lost.

Notice that satisfying **(2)** entails that content is causally potent. For, intervention on the structural similarity in **(1)** *just are* interventions on what V represents - that is, its *contents*. But, as seen above, these interventions also modify the agent's odds of success: the better the similarity, the better the agent's odds. This is enough to make V's content causally potent under an interventionist notion of causality (Gładziejewski & Miłkowski 2017): changes in V's contents *cause* an agent to be more likely to non-accidentally succeed or fail.

Here, I wish to highlight two ways in which the structural similarity between V and T can be worsened - and so, two ways to non-accidentally decrease an agent's odds of success. First, the similarity can be worsened because single vehicle constituents of V map onto *many* target constituents of T. This is one way to violate **(a)**. I will call it an *(a)-violation*. Secondly, the similarity between V and T may be degraded because two constituents of V display the corresponding constituent of T as being in a relation that does not in fact hold. This is one way of violating **(b)** - and I will call it a *(b)-violation*. Resorting to the map example may help clarify both cases. When an (a)-violation occurs, one bit of the map "stands for" multiple bits of the terrain - like a dot on a map representing *both* Paris and Rome. When a (b)-violation occurs, the map inaccurately displays the terrain by displaying certain parts of it being in a relation that does not in fact hold between them - like a map displaying Rome north of Paris. There are of course *further* ways in which the structural similarity between V and T may be worsened: **(a)** and **(b)** can be violated in many other ways. But my arguments won't hinge on *these* violations, so I won't discuss them.

Point **(3)** mentions decouplability. Decouplability is an essential feature of all representations, which captures the idea that representations represent their target even when their target is not causally affecting them or the agents relying on them (cf. Orlandi 2020). A map can be used even when the mapped terrain is not causally interacting with the map or its user: for example, a map of Tokyo represents Tokyo even if it, and its user, are located in Buenos Aires. Minimally, then, decouplability can be unpacked as follows: V is decoupled from T when T is not causally influencing V - for example, by causing its tokening (cf. Gładziejewski 2015; 2016). Notice, however, that **(3)** requires something *more* than decouplability thus spelled: it requires decouplable representations to still play the action guiding role specified by **(2)** when decoupled. So, for a map of Tokyo to fully satisfy **(3)** it is not enough that it continues to represent Tokyo while located in Buenos Aires. It must also perform its action guiding duties while in Buenos Aires - for example, by

[8] This *Ceteris paribus* clause is meant to exclude cases in which excessive degrees of similarity stand in the way of representational usage, as in the case of an hypothetical map in 1:1 scale.

allowing the map user to plan her trip to Tokyo in a way such that the plan's odds of non-accidental success depend on the degree of similarity holding between the map and Tokyo.

Lastly, **(4)** is entailed by **(2)**[9]: if V guides S's actions in regards to T as required by **(2)**, then the degree of similarity between V and T is reflected in S's odds of success. Hence non-accidental behavioral successes can act as reliable (through defeasible) indicators of representational accuracy: pragmatic successes indicate representational successes, and pragmatic failures indicate representational failures - thereby allowing the detection of representational errors. (cf. Gładziejewski 2015; 2016, see also Bielecka and Miłkowski 2020 for further elaboration).

Summing up: structural representations are representational vehicles **(1)** structurally similar to a target, **(2)** whose structural similarity guides an agent's action aimed at that target, **(3)** that can do so even when decoupled from their target and **(4)** that allow their user to determine their representational accuracy *via* the success-rate of the actions they guide. NSRs are just structural representations realized in the neural medium. Thus, if they are present, we should be able to observe NSRVs: neural vehicles satisfying **(1)**-**(4)**.

But, does our neuronal activity *really* realize such vehicles? I think the existing neuroscientific data motivate a negative answer.

3 - Are bona fide neural vehicles vehicles of neural structural representations?

To determine whether neural vehicles satisfy **(1)**-**(4)**, one must first determine what neural vehicles are. Here, I take neural vehicles to be *neuronal responses*, which I analyze at three distinct spatiotemporal levels: the level of individual neuronal responses (**§3.1**), the level of neural maps (**§3.2**), and the level of entire activation spaces (**§3.3**). In all cases I claim that they do not, and, indeed, *cannot*, satisfy **(1)**-**(4)**.

What justifies this focus? Bluntly, the fact that neuronal responses are most often considered the relevant representational vehicles upon which neurocognitive processes operate (see, for example, Friston 2005; Mesualm 2008; Backer *et al*. 2022; Frisby *et al*. 2023). They're the vehicles cognitive neuroscience focuses on the most - the ones that are most central to its explanatory practices. However, it should be noted that neuronal responses are not the *only* representational vehicles cognitive neuroscience deals with. So, I will briefly consider some other alternative neural vehicles, claiming that they do not qualify NSRVs either (**§3.4**). A brief summary of the entire discussion will then close this whole section (**§3.5**).

3.1 - Individual neuronal responses are not vehicles of neural structural representations

Individual neurons respond to stimuli selectively: different stimuli elicit different responses. Typically, neurons have one *preferred stimulus*, which elicit the strongest response. Preferred stimuli vary depending from neuron to neuron, reflecting their specialized roles. For example, neurons in the primary visual cortices respond to simple visual stimuli like oriented bars (cf. Hubel

[9] At least, in sufficiently complex systems: we surely could design a robot whose central control system allows the tokening of states satisfying **(1)**-**(3)** but not **(4)**. However, since the paper focuses on brains (and brains are arguably *sufficiently complex*) I will somewhat critically take **(2)** to entail **(4)**.

& Wiesel 1968). Neurons in hierarchically higher layers of the visual cortex respond to more complex stimuli - for example, neurons in area MT respond to movement directions (cf. De Angelis & Newsome 1999). Neurons further away from sensory areas respond to even more complex stimuli (or features thereof): the parietal cortex houses neurons responding to specific quantities (Nieder *at al*. 2006), the inferior premotor areas of the frontal cortex house neurons that respond to specific actions (Kohler *et al*. 2002) and, apparently, there are even neurons in the inferior temporal cortex preferring specific individuals (Quiroga *et al* 2005). Thus, individual neurons have preferred stimuli of different sorts, which they are often said to *represent*. But are these representations NSRs? Are they underpinned by NSRVs?

It is a bit hard to provide a direct answer to these questions. Sure, NSRVs should be observable and manipulable as any other component of a mechanism - but this time it is a bit unclear what we should be looking at (or thinking with) *exactly*. For, "individual neuronal response" can be read in at least three different ways: (i) as designating individual spikes (i.e. single neuronal discharges), (ii) as designating spike trains (i.e. sequence of discharges) and (iii) as designating a neuron's firing rate compared to a baseline. Options (i)-(iii) all pick up a *bona fide* representational vehicle supporting a specific representational scheme (see, for example, Dayan and Abbott 2005; Brette 2015). Thus, the claim that individual neuronal responses are NSRVs can be read in at least three different ways. As a consequence, it is not immediately clear what sort of observations and manipulations would support it.[10]

Now, whilst interpretations (i)-(iii) are all possible, I want to suggest that they all face certain important challenges, whose collective weight seems enough to reject the idea that individual neuronal responses may qualify asNSRVs under *any* interpretation.

First, it is very hard to see how an individual neuronal response could structurally resemble its target - be it an oriented bar or an individual person. This is because it is very hard to see how the vehicle (i.e. the individual response, however interpreted) could be non-arbitrarily decomposed into vehicle constituents as requested by (a). It is not at all clear what could count as a vehicle constituent of a single neuronal response: a "part" of a spike, an individual spike (or sequence of spikes) in a spike train, part of the voltage emitted, a fraction of the firing rate, part of the neurotransmitters emitted, or something else entirely? All these options pick up certain bona fide parts of a single neuronal response. Yet, there seems to be no privileged way to choose between them (cf. Maley 2023): the choice of vehicle constituents seems entirely arbitrary. This is a serious problem when it comes to satisfying (1). Of course, I don't want to deny that we may *discover* that there are functionally relevant, non-arbitrary ways to decompose individual neuronal responses. But we've not discovered them yet. So, even supposing that one such partition exists (which is something my dialectical adversaries should argue for!) we've not yet observed the relevant NSRVs, for we simply do not know what that partition is. Moreover, even if a privileged, non-arbitrary way to identify vehicle-constituents in individual neuronal responses were to be found, we still would have to specify what sort of relevant *relation* holding amongst the vehicle-constituents as specified by **(b)**. A task as daunting as the previous one.

[10] Notice that the claims that neuronal maps and activations spaces are vehicles of NSRs are not similarly ambiguous: both claims express a form of population coding, which is a special case of *rate* coding. No interpretation of these claims in terms of single spike trains (or single spikes) is possible.

Secondly but not least importantly, such tasks are not just daunting. They are also entirely unmotivated - at least insofar the explanatory practices of present day cognitive neuroscience go. For, whilst contemporary cognitive neuroscientists typically assume that individual neuronal responses represent individual targets, they do *not* claim that specific *parts* of neuronal responses represent specific *parts* of the target, nor do they claim that relations between parts of neuronal responses represent relations between parts of the target. But that's exactly the way in which structural representations represent. Moreover, I suspect that claims such as "The first spike of the spike train represents the leftmost bit of the oriented bar" or "the fact that spike v_a preceded spike v_b represents the fact that a part t_a of the oriented bar is left of a part t_b of the same bar" would be considered not just unjustified, but entirely *exotic* by the majority of cognitive neuroscientists. So exotic, indeed, to be a *bona fide reductio* of the idea that individual neuronal responses are NSRVs.[11]

Summing up: the claim that individual responses are NSRvs is hard to "cash out", it yields extremely exotic conclusions and it is entirely unjustified by the current practice of cognitive neuroscience. Individual neuronal responses are in fact typically described as "indicator" or "detector" representations (cf. Ramsey 2003; Williams & Colling 2017; Gładziejewski & Miłkowski 2017; Backer *et al*. 2022).[12] On this view, the firing of a neuron does not provide an inner model of a target which replicates the target's inner structure. Rather, the firing of a neuron simply signals the presence of the target at the time of firing. So, the actual practice of cognitive neuroscience - that is, the observations and manipulations that cognitive scientists actually carry out - does not suggest or motivate the claim that individual neuronal responses are NSRVs. If anything, individual neuronal responses are said to be the *vehicle constituents* of individual structural representations (cf. Williams & Colling 2017; Gładziejewski & Miłkowski 2017) - a view whose two different popular incarnations will be discussed in (**§§ 3.2** and **3.3**)

Before doing so, however, I wish to discuss an increasingly popular line of argument that purportedly demonstrates that individual neuronal responses are NSRs precisely *because* they are indicators. To anticipate: I will claim that indicators *cannot* be structural representations, as indicators can *never* satisfy **(2)** and **(3)** in conjunction. To keep things in good order, I'll do so in a separate subsection. Readers more interested in hands-down philosophy of neuroscience might wish to skip to (**§3.2**). Readers more interested in the "indicators *vs* structural representations" debate are instead encouraged to read on.

3.1.1 - Why indicators cannot be structural representations (and individual neuronal responses can't be neural structural representations)

According to some, indication is a *special case* of structural similarity (Nirshberg & Shapiro 2020) and indicators are a special case of structural representations (Morgan 2014; Facchin 2021b). Consequently, *neural* indicators are a special case of NSRs. Were this line of argument correct, the fact that we have observed individual neuronal responses (in one of the readings of the term) being

[11] One could still argue that individual neuronal responses represent what they represent *because* they are part of a larger structural representation. Notice, however, that, in such a case, individual neuronal responses would not be NSRVs, but only *vehicle constituents* of a larger NSRV. At any rate, **§§ 3.2-3.4** will consider putatively larger vehicles, concluding that they don't qualify as NSRVs either.

[12] Piccinini (2020a) might, under a certain reading, be an exception - but he really seems more concerned with *populations* of neurons rather than individual neurons. I will thus deal with his view in (**§3.2**).

indicators is sufficient to establish the fact that we have observed individual neuronal responses being NSRVs.

The argument claiming that indicators are a special case of structural representations goes roughly as follows. First, notice that there is a one-to-one correspondence between indicator states and indicated target: for example, each possible height of a thermometer's mercury ball maps onto one temperature. Thus, **(a)** obtains. Notice that there is always a (indicator specific) relation such that **(b)** obtains: if the mercury bar height v_a is *higher than* v_b (i.e. $v_a R v_b$) then the temperature t_a is *hotter than* t_b (i.e. $t_a R^* t_b$). Minimally - and most essentially, as Facchin (2021b) points out - temporal relations between corresponding indicator and target states must satisfy **(b)**: if v_a is *present n seconds after* v_b (i.e. $v_a R v_b$), then the temperature t_a *followed* t_b *after n seconds* (i.e. $t_a R t_b$); else, the device would not be indicating the temperature in the first place. Thus, all indicators satisfy **(1)**. They also satisfy **(2)**: if a system relies on indicators to organize its behavior, then the better its indicators indicate, the better the system's odds. But, since indication *just is* a structural similarity, the better the structural similarity, the better the odds - exactly as **(2)** requires. What, then, about **(3)** and **(4)**? Here, there seems to be no common argument. Morgan (2014) does not discuss them, and Facchin (2021b) only presents certain examples suggesting that *some* (fairly complex) indicators can satisfy them. Are individual neuronal responses amongst these indicators?

No, they are not. They fail to satisfy both **(3)** and **(4)**. Consider **(4)** first, as it poses the smallest problem. As noted above (**§2**), **(4)** is entailed by **(2)**: if V guides S's actions about T, then the failure (or success) of these actions indicates whether V is an accurate representation of T. Yet, many indicators - especially these in the primary sensory or motor cortices - do not indicate the targets of our actions. For example we rarely (if ever) act on straight bars and the other stimuli the neurons in our primary visual cortex indicate (cf. Hubel & Wiesel 1968). So, these neurons - or better, their responses - fail to satisfy **(4)**. So, it can immediately be concluded that at least some individual neuronal states are indicators but not NSRs. One, however, could perhaps solve this problem noticing that whilst what these indicators indicate is not the target of our actions, it is nevertheless *part* of the targets of our actions, in a way that coil allow us to assess the semantic status of these receptors too (thanks to XYZ for having noticed this). So, the problem with **(4)** is not fatal - or at least, not fatal as I'd like it to be.

Luckily for me, the problem with **(3)** has the desired dose of theoretical lethality. Could a neural indicator state be used offline - in a way that is decoupled from its target? The answer seems positive: we have compelling evidence that "offline" and "online" cognition rely on the same neuronal resources (e.g. Albers *et al*. 2013). If those resources include individual neuronal responses (and they do) and these responses are due to indicators (and they are), then indicators are used "offline", in absence of the causal touch of the represented target. *But if so, then* **(2)** *actually fails to obtain*. To see why this is the case, recall, as Facchin (2022b) stressed, that the indicator-target structural similarity is essentially based upon certain *temporal* relations holding amongst indicator states and target states. If v_a *follows* v_b *after n seconds* (i.e. $v_a R v_b$), then t_a *follows* t_b *after n seconds* (i.e. $t_a R t_b$) - else, V would simply fail to indicate T. But when V is used offline, in a decoupled manner, the temporal succession of indicator states and the temporal succession of

indicated states *must* diverge - else, V would simply be indicating T online.[13] Thus, when an indicator V is used offline, if ($v_a R v_b$), then not ($t_a R t_b$). But this is just a (b)-violation of the relevant structural similarity: to use V offline, one represents certain relevant states of T in temporal relations that in fact do not hold amongst them. If **(2)** were the case, that (b)-violation would hinder S's behavior, making S more likely to non-accidentally fail. But that's not the case. For, the ability to re-use one's neuronal resources to cognize offline has clearly a high adaptive value. It allows an agent to test behavioral strategies offline, without suffering the consequences of real, "online" failure (cf. Dennett 1996). It also allows an agent to anticipate various environmental challenges, so as to take action before the nefarious consequences of such challenges unfold (cf. Pezzulo 2008). So, the offline usage of indicators leads the agent to more frequently and more robustly achieve behavioral successes, it *increases the* subject's odds of success. Thus, if **(3)** obtains, then **(2)** fails to obtain. *And vice versa*: when (for all indicator and target states) v_b follows v_a after *n* seconds and t_b follows t_a after *n* seconds, **(2)** obtains, but, as seen above, the indicator is not decoupled. So, when **(2)** obtains **(3)** fails to obtain. In summary: for all neural indicators **(2)** and **(3)** *never* obtain together - and so neural indicators are not NSRs. Worse still, the argument can be easily generalized to *all* indicators: hence - *pace* (Morgan 2014; Facchin 2021b) - *no* indicator qualifies as a structural representation.

Whilst the two arguments provided above are already sufficient to conclude that neural indicators are not NSRs (and, more generally, that indicators are not structural representations), I wish to point out a further problem in (Morgan 2014; Nirshberg & Shapiro 2020; Facchin 2021b). Thus far, I've conceded that these arguments are sufficient to show that indicators satisfy **(1)** and **(2)**. Now, I wish to point out that (at least in this context) they do not actually show that **(1)** and **(2)** obtain. More precisely, I want to claim that these arguments do not show that *individual neuronal responses* (under any interpretation (i)-(iii)) are structurally similar to their targets; let alone that that structural similarity plays a relevant action guiding role. For, the structural similarity these arguments individuate is built around *numerous individual indicator states* and certain relations holding amongst them. So, these arguments do *not* show that *individual indicator states* are structurally similar to the targets they individually indicate. These arguments only show that an indicator's entire *range of states* is structurally similar to the indicator's *range* of targets. Applied to neurons, then, these arguments do *not* show that individual neuronal responses are structurally similar to their targets; only that a neuron's response profile[14] is structurally similar to the entire range of stimuli eliciting an activation of that neuron.

Here's another way to make the same point: consider how Morgan (2014) Nirshberg & Shapiro (2020) and Facchin (2021b) claim that the conditional in **(b)** (i.e. ($v_a R v_b$) → ($t_a R^* t_b$)) does in fact

[13] The reason for this is simple: if v_b follows v_a after *n* seconds and t_b follows t_a after *n* seconds, then v_b and t_b are co-occurring and so V is actually just indicating the state of T. A toy example to ease the understanding of this point: Suppose I have a "Jennifer Anyston" neuron (Quiroga *et al*. 2005), that is, a neuron in the task of indicating the presence of Jenifer Anyston. Suppose I detected Jennifer Anyston at t_1, and call the activated state of my detector v_a. Now, let some time pass, and suppose my Jennifer Anyston detector re-activates, entering a detector state v_b at time t_2. Let *R* be the temporal relation between v_a and v_b; e.g. v_a is followed by v_b after *n* minutes. So, $v_a R v_b$ holds. Does $t_a R t_b$ hold too? If so Jenifer Anyston was actually present *n* minutes after I first detected her. So, the second detector state v_b is clearly *not* decoupled from Jenifer Aniston - I'm actually detecting her at t_2 too! So, in order for v_b to be used offline, it must be the case that I'm not actually detecting Jennifer Anyston, and so $t_a R t_b$ must fail to hold.

[14] Or - to anticipate a theme from (**§3.3**) - a neuron's "activation space"

hold. They claim they hold, for example, because in indicators, if v_a follows v_b after n seconds (i.e. v_aRv_b), then t_a follows t_b after n seconds (i.e. t_aRt_b). Now, whilst this claim is true, they are quite obviously treating individual indicator states as *vehicle constituents*, rather than vehicles. Applied to neurons, this means that they are not treating individual neuronal states as vehicles, but *only* as vehicle constituents. But what is at stake in this paper is whether individual neuronal responses are NSRVs, not whether they are vehicle constituents of larger NSRV.

So, whilst the arguments by Morgan (2014) Nirshberg & Shapiro (2020) and Facchin (2021b) *do* show that indicators and their targets are structurally similar (in a quite specific sense), using their arguments to claim that *individual neuronal responses* are structurally similar to their targets is, if not an entirely unwarranted move, at least a move that "smuggles in" a significant change in the focus of the analysis. And, at any rate, the argument they offer prevents *individual neural responses* from being NSRVs: they are, at best, vehicle constituents.[15]

So much so for the idea that individual neuronal responses may qualify as NSRVs. But about other *bona fide* neural vehicles?

3.2 - Neural maps

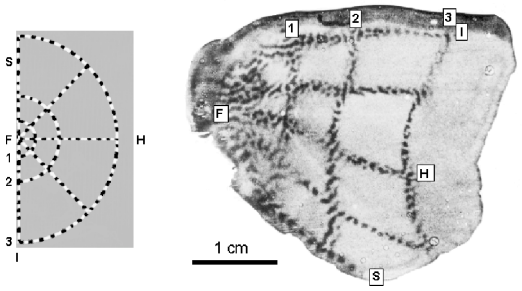
Above, I've argued that individual neuronal responses are not NSRVs. But what about the responses of *multiple* neurons?

Piccinini (2020a), argues at length that various types of cortical maps - including the retinotopic map in the primary visual cortex and the motor and sensory *homunculi* - qualify as NSRVs. Ramsey (2016), Shea (2018) and Gładziejewski & Miłkowski (2017) all claim that certain neurons in the hippocampus of rats are connected in a map-like way, so as to structurally represent the rat's environment.[16] So, many authors suggest that the real NSRVs are responses of *multiple* neurons organized in a map-like way.

These arguments can call upon a wealth of well-known neurophysiological and neuropsychological data. For example, Piccinini (2020a, p. 271) stresses the retinotopic organization of the primary visual cortices (V1), nicely displayed in **figure 1**:

[15] This isn't however, to deny that the arguments by Morgan, Nirshberg & Shapiro and Facchin show something important; namely that both indicator and structural representations are *analog* in an important sense of the term, and thus that both indicator and structural representations are *analog* representations in one relevant sense of the term (cf. Maley 2021a; Lee *et al*. 2022 for further discussion).

[16] See (Thomson and Piccinini 2018; Bechtel 2008; 2014) for a non NSRs-centric representational account of these neural structures.

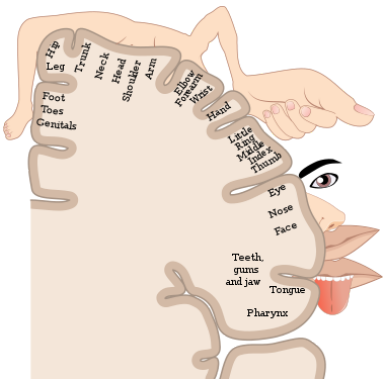


Caption - Figure 1: Cortical topography of V1: the spatial structure of the stimulus (left) is mirrored - in a systematically distorted fashion - by V1 neurons (right). The same topological structure, however, is instantiated in both. [source: The development of topography in the visual cortex: a review of models. N. Swindale - PERMISSIONS STILL TO BE ASKED]

The neurons constituting V1 them are spatially organized so as to replicate (a tweaked version of) the spatial structure of the original visual stimulus (cf. Tootell *et al*. 1988). If neuron v_a is *left of* neuron v_b, then t_a (i.e. whatever v_a is responding to) is left of t_b. This is a clear structural similarity tying together the neural map and its representational target. Further, Piccinini stresses that the columnar organization of V1 contains many "smaller scale" cortical maps representing significant properties:

> "V1 contains multiple fine-grained topographically organized feature maps of such properties embedded in the larger-scale retinotopic representation of space. For instance, those neurons selective for horizontally oriented bars tend to cluster together in cortical columns in V1, and nearby columns contain neurons that are tuned to similar orientations" (Piccinini 2020a, p. 272).

So, if column v_a is close to column v_b, then t_a is similar to t_b. Similar "smaller scale" maps are found in many neural areas. For example the neurons area MT (a further visual area particularly sensitive to movement) are arranged so as to compose a "movement map". Neurons that prefer similar direction of motion cluster into columns, and columns are spatially organized so that spatially close columns prefer similar movements (cf. De Angelis & Newsome 1999). The closer two columns (or two neurons) are, the more similar the velocities they respond: if v_a is close to v_b. then t_a is similar to t_b. More intuitively strikingly still, there are the cortical "homunculi" and "simunculi" drawn by Penfied and Woolsey (cf. Penfield and Brodley 1937; Woolsey *et al*. 1952). It's hard to look at them without noticing how nicely the spatial organization of these neurons "recapitulates" the spatial organization of bodily parts - for one example, see **figure 2**.



Caption - Figure 2: The sensory homunculus. Note how the spatial relations between the cortical areas "mirror" the spatial relations between the represented body parts Source: Wikimedia Commons (https://commons.wikimedia.org/wiki/File:Sensory_Homunculus-en.svg).

Notice how easily the relevant structural similarity can be *seen* in **figure 1** and **figure 2**. Isn't it simply obvious that these structures *are* structurally similar to their target, in a way that clearly satisfies **(1)**?

Whilst these structures obviously *seem* structurally similar to their targets, it is not at all obvious that they *are* - or so I will later argue. But before doing so, I wish to notice that even if such similarities were present, the fact that their presence is obvious *to us* does not entail their presence is at all obvious to our neurocognitive mechanisms - indeed, it seems that our neurocognitive mechanisms are blind such similarities in the execution of their tasks. And, for this reason, these similarities fail to satisfy **(2)**. Consider, for example, the somatotopic organization of the cortical homunculi - and the structural similarity it underpins. Does the somatotopic *spatial arrangement* of these neurons guide our actions as required by **(2)**? *Prima facie*, the answer is negative. Imaginary interventions that modify *only* the somatotopic organization of the homunculi (i.e. the relative spatial locations of the neurons constituting it) do not seem to have any effect on our behavior. After all, if they modify *only* the somatotopic organization of these neurons, they leave intact their input-output profile and mutual connections, allowing the homunculus they constitute to contribute to an agent's behavior *in the same way* in which a somatotopically non-modified homunculus would. Changes in the somatotopicity of homunculi - and the structural similarity they underpin - do neither increase nor decrease the agent's chance of success. So, **(2)** fails to obtain.

One could object that similar though experiments are ill-suited to determine whether **(2)** obtains or not. Looking at some *real* experiments, however, yields the same verdict. Consider, for example, the data collected by Hartmann *et al*. (2016).[17] Simplifying to the extreme, they equipped rats with prosthesis enabling them to perceive and respond to infrared lights. The prosthesis were "caps" of infrared sensors (allowing for a 360° panoramic infrared vision) that communicated with the rat's "sensory homunculus" (i.e. their primary somatosensory cortices). Crucially, they could do so in a way that either respected or flouted (to various degree) the homunculus's somatotopic organization - e.g. the front-facing infrared sensor could be connected with the head of the rat's homunculus

[17] Though it should be noted that the experimental interventions in (Hartmann *et al*. 2016) are not interventions *only* on somatotopicity, as they always also change the artificial sensors from which neurons receive inputs. Here, I will ignore this complication for the sake of simplicity.

(respecting somatotopicity) or with its rear or side (flouting somatotopicity). Now, Hartmann and colleagues report that all rats managed to achieve a high success rate in the experimental task (infrared light discrimination), regardless of the degree of somatotopicity of their prosthesis. Sure, the better the somatotopicity, the *faster* behavioral success came. But, eventually, even rats equipped with "non-somatotopic" prosthesis were eventually able to perform at the level of rats equipped with "somatotopic" prosthesis. This clearly violates **(2)**, according to which the degree of somatotopicity should be reflected in higher or lower odds of behavioral success.

Now, one could object that these data are less clear cut that I'm making them appear - after all, rats with "non-somatotopic" prosthesis learned how to face the experimental task more slowly than rats with "somatotopic" prosthesis, and this might be counted as one way in which degrees of somatotopicity influence the agent's odds of non-accidental success during the learning phase. I'm not persuaded that this is the case (why should the degree of somatotopicity matter only during the learning phase of a task?) - but even if it were the case, other experimental data on *homunculi* can be marshaled to support my conclusion. For example, Chakrabarty & Martin (2000) have found that, during postnatal development of the primary motor cortex (i.e. the motor homunculus) the number of sites representing more than one limb *increases*. This suggests that such "multi target" sites are needed to effectively control movements - something that improves during postnatal development. And yet, "multi target" neurons clearly degrade the structural similarity in **(1)**, as they are a case of an (a)-violation (cf **§2**).[18] So, a *worsening* of the structural similarity correlates with an *increase* of performance, blatantly violating **(2)**. Martin *et al* (2005) present similar data, suggesting that increases in "multi target" neurons are positively correlated with increases of motor expertise.[19]

The evidence above gestures towards a point that can perhaps be less messily expressed (and generalized beyond homunculi) as follows. The structural similarity of cortical maps is based on certain *spatial* relations holding amongst the map's constituents - that is, spatial relation between neurons. Now, according to a standard neuroscientific picture, neurons and neural maps contribute to an agent's behavior in virtue of their information-signaling properties; roughly, their input-output behavior. Their input output behavior is determined by a number of features, including a neuron's sensitivity to stimuli, their baseline firing rate, the connections they have with other neurons and the nature of such connections (excitatory or inhibitory) and other features. *Spatial* features, however, do not influence their input-output behavior. So, they don't contribute to an agent's behavior. Hence they can be varied *ad libitum*, creating arbitrarily large (b)-violations, without influencing an agent's behavior and its odds of success. And this, of course, means that they do not play the action guiding role required by **(2)**.

Notice that I'm *not* claiming that the topographic organization of cortical maps does not play any relevant functional role. Not all functional roles of neuronal structures must be representational or cognitive (Haueis 2018). Perhaps the topographic organization of cortical areas minimizes wiring,

[18] More on this point below.

[19] One could object that motor homunculus is not a good example, because it is not at all clear how the primary motor cortex represents our body and its movements (cf. Thomson and Piccinini 2018; Piccinini 2020a). This, however, is more a problem for the defender of NSRs than for me: how can we claim that the motor homunculus is a NSR if we do not know what it is structurally similar to?

speeding up neural signaling (cf. Blauch *et al*. 2022).[20] Maybe it reduces metabolic costs (cf. Sterling and Laughlin 2015). Or perhaps it is just a side effect of certain relevant evolutionary or developmental constraints - or maybe it is due to all three, and perhaps even more, factors simultaneously (Cf. Graziano & Aflalo 2007, p. 239). I'm not denying these (or similar) claims. I'm only denying that the topological organization plays the representational role **(2)** captures. This is entirely compatible with it playing *other* biological - or even cognitive - roles (cf. Graziano 2011). To deny a car's brakes makes it accelerate is not to say brakes are useless!

One could retort that the argument above is not fully general. In the case of the spatial map in the rat hippocampus, for example, what matters are not the *spatial* relations amongst neurons, but rather the relation of *inducing activation*. If neuron v_a tends to induce the activation of v_b, then t_a is close in space to t_b (cf. Moser *et al*. 2008). But this relation is a *functional* relation, the changing of which changes the way in which inputs are turned into outputs. Hence **(2)** seems to obtain, and the argument provided above does not apply. And, perhaps, some similar functional relation might similarly rescue the neural maps discussed above. For example, the motor homunculus might not underpin a NSR of our body, but rather a NSR of our action (cf. Graziano 2016). If that were the case, my focus on somatotopicity might just have distracted from some *other* functionally relevant structural similarity.

Even if that were the case, however, there would still be a significant problem. In general, neurons (including the neurons of cortical maps) do not respond to just *one* stimulus. Sure, they respond most strongly to their preferred stimuli, but it makes sense to say that neurons have *preferred* stimuli only because they respond to many different stimuli. Moreover, the response profile of neurons is typically influenced by *multiple* parameters of a stimulus. For example, MT neurons are not just sensitive to motion direction, but also the retinal position of the stimulus, its size, the speed of motion and its binocular disparity (Born and Bradley 2005, P. 164). Hippocampal place cells do not respond *only* to place, but also to odors, tactile inputs, recognizable chunks of experiences, and the relative timing of certain events (Wood et al. 1999, 2000; Itskov*et al*. 2011; Kraus et al. 2013; Sun et al. 2020). Even the neuronal cells constituting the "cortical homunculi", probably the most well known and the most intuitively compelling NSRs, do not always code for single bodily parts (see Penfield and Brodley 1937; Penfield and Rasmussen 1957; Woolsey *et al*. 1952; Kwan *et al*. 1978; Wasserman *et al*. 1992; Schieber 2001; Aflalo & Graziano 2006). Indeed, some neurons of the "motor homunculus" appear to code (and control) complex whole-body configurations, in a way that clearly stands in the way of **(1)** (Gordon *et al*. 2022): if these neurons are vehicle-constituents of the NSRV representing our whole body, they *can't* be representing our whole body without violating **(1)**! All these are significant and systematic (a)-violations of the relevant structural similarity. So, in general, the neat one-to-one mapping from discrete and well-identified "bits" of the neural map to discrete and well-identified bits of the world is a *huge* idealization of the neurobiological reality.[21] As far as neuroscience shows us, (a)-violations are the

[20] Though others suggest that wiring length minimization does not strongly correlate with topographic organization (cf Yarrow *et al*. 2014).

[21] Penfield was explicit on this point. He considered his homunculus as "a cartoon of representation in which scientific accuracy is impossible" intended to be used as an "aid to memory" (both quotes from Penfield and Rasmussen 1950, p.56)

rule, not the exception, in cortical maps. So it seems that, as a general rule, **(a)** fails to obtain. *A fortiori*, **(1)** does not obtain too.[22]

One could claim that these data pose no threat to **(1)**, as they only show that NSR are much messier than textbook philosophical examples lead us to suppose (thanks to XYS for this objection). But these data do not "just" complicate the picture. They complicate the picture in a way that directly threatens the obtaining of **(1)** by showing that the relevant vehicle constituents do *not* map onto target constituents in the desired manner. They *don't* show that **(1)** obtains, but in a much messier manner than textbook examples indicate. They show that **(1)** does not obtain.

One could further claim that these data pose no threat to **(1)** because structural similarities between vehicles and targets need not be perfectly accurate nor total. Imperfect, partial, distortive similarities are sufficient to satisfy **(1)** too (cf. Williams and Colling 2017; Shea 2018, pp. 140-142). And indeed, sometimes *distortive* similarities might be more functional than non-distortive ones: think the way in which maps of underground metros are way more readable when they do *not* display the actual distance holding between the various metro stations. I think this is an important claim that gets something importantly right.. However, I still think that, in the present context, it is insufficient to rescue **(1)**.

For, appealing to approximate similarities allows **(1)** to tolerate *local* (a)-violations and/or (b)-violations, *global* (a)- and/or (b)-violations can't be tolerated. A map can tolerate a (a)-violation (e.g. representing Rome and Paris with a single point) only if it correctly represents other places (e.g. because it represents Lyon and Florence as two distinct points, the former north of the latter). Else, it ceases to be a map in any recognizable sense. And the same goes for (b)-violation. Thus, (a)- and (b)-violations cannot be global. In the case at hand, however, the (a)-violation seems to be if not global at least *extremely* widespread. Neurons responding (and mapping to) single targets, if they exist, are rare exceptions - so rare, indeed, they're yet to be found.

But perhaps one could argue that, unlike cartographic maps, cortical maps might tolerate *global* (a)- and/or (b)- violations. After all, neural representations have unique properties, and public representation offers only a limited, and mostly analogical, guidance to the understanding of neural representations (thanks again to XZY for this objection). Whilst this objection, if successful, would rescue **(1)**, I'm not entirely sure that it makes sense; and I think that even if it were sensical, it could not be accepted.

I'm not sure that the objection is sensical because I'm not sure that there is a real difference between something that satisfies **(1)** while allowing for systematic (a)- and/or (b)-violations and something that simply *fails to satisfy* **(1)**. I really don't have the faintest idea of how that difference could be spelled out and articulated - *prima facie*, something allowing for systemic (a)- and/or (b)-violations is simply something that does *not* satisfy **(1)**. If there is a difference between the two, I challenge the objectior to spell it out in a clear manner.

Now suppose, for the sake of discussion, that such a difference has been spelled out in a way that is sufficient to rescue **(1)** - that is, concede to the objector that **(1)** obtains. Would this lead the

[22] As an aside, notice that the same state of affairs prevents us from considering these neurons and neuronal regions *indicators* in any straightforward and intuitive way.

objector to win the day? I don't think so. For, since **(a)** and **(b)** partially determine the semantic properties of structural representations, their global violation yields degenerate semantic properties. And these degenerate semantic properties seem to impede cortical maps to be counted as NSRs, for the possession of such degenerate semantic properties is incompatible with the semantic transparency that characterizes structural representations. Further, such degenerate semantic properties make cortical maps unable to play the causal role that characterizes structural representations. So, we can't really *coherently* accept that NSRV can allow for systematic (a)- and/or (b)-violations. Let me elaborate on this (I fear a bit clunkily written) passage.

Recall (**§2**): in structural representations each vehicle constituent represent the target constituent onto which it maps, and the fact that vehicle constituents stand in certain relations represent the fact that corresponding target constituents stand in corresponding relations: $v_a R v_b$ represents that $t_a R^* t_b$. But actual neural responses "map onto" more than one target constituent - neurons do not respond *only* to their preferred stimuli. So, in the case at hand, v_a does not map only onto t_a, it maps also onto t_{a^*}. But then, what does $v_a R v_b$ represent? $t_a R^* t_b$, $t_{a^*} R^* t_b$ or some mixture of the two? I want to argue that *no* answer to this question can be accepted by a defender of NSRs. For, each answer fails to deliver contents with the requested semantic transparency. Moreover, since the content lacks the desired semantic transparency, it is unclear when V represents T. Hence it is unclear whether V is able to play the action guiding role imposed by **(2)**. Since **(4)** is entailed by **(2)**, **(4)** is in danger too.

To see why this is the case, suppose, first, that $v_a R v_b$ represents only $t_a R^* t_b$ - or only $t_{a^*} R^* t_b$. Notice that this is a fairly substantial supposition: it amounts to supposing that V actually satisfies **(1)** and so has the required semantic transparency. But even with a substantial supposition in place, it is not yet determined whether $t_a R^* t_b$ or $t_{a^*} R^* t_b$. The supposition is that only one of the two is represented - but now it is necessary to determine *which one* is represented. But *how* could we do that? What V represents is determined by the relevant structural similarity V bears to some target - but that structural similarity *does not* discriminate between $t_a R^* t_b$ and $t_{a^*} R^* t_b$. So, V's content is indeterminate: Sure, V represents one and only one target T, but *which individual target T is represented is left entirely open*.[23] V is thus semantically transparent in name only. Further, since T is indeterminate, whether **(2)** and **(4)** obtain is left entirely unclear. If we don't know what V is structurally similar to, we can't determine whether increasing (or decreasing) that similarity

[23] A tempting and obvious solution to this problem is that of resorting to a form of informational (or information-based) semantics; that is, claiming that each neuron "maps onto" the stimulus about which it carries the most information (cf. Wiese 2017, pp. 219-223). However, such informational linkages are not only far from trivial to ascertain (cf Brette 2019), they also seem unable to ascribe determined contents (Artiga & Sebastian 2018; Rosche & Sober 2019) - and, more generally, theories of structural representations interact poorly with informational accounts of content (cf. Facchin 2021a). A second solution is that of appealing to the agent's actual context. But this solution can only work in *some* cases of successful online behavior. If the relevant vehicle is used in a decoupled manner, in service of offline cognition, then there is *nothing* in the agent's context that can discriminate between $t_a R^* t_b$ and $t_{a^*} R^* t_b$ - else, the agent' would not be decoupled from at least one of them. Moreover, the solution fails to cover for unsuccessful online behaviors (how can context in principle determine whether the subject is misrepresenting the actually present $t_a R^* t_b$ or whether it is correctly representing an irrelevant or absent $t_{a^*} R^* t_b$? after all, both options make the agent's fail in a given context) and successful online behaviors in which both $t_a R^* t_b$ and $t_{a^*} R^* t_b$ are present (context does not discriminate between the two). So, the solution does not generalize and fails to appropriately restore content determinacy. Other solutions are far less obvious, and thus cannot be considered here.

increases (or decreases) the agent's chance of success. V would thus be a vehicle of a structural representation in name only. Moreover: the fact that neurons commit systematic (a)-violations is functionally relevant - it improves the way in which our neurocognitive mechanisms work (Chakrabarty & Martin 2000; Martin *et al*. 2005). If the way in which such mechanisms function really is best explained representationally, a representational explanation should should be expected to *emphasize* that fact, rather than hiding it under the carpet assigning these representational vehicles a single representational content *by fiat*.

So, in the case at hand, a representational explanation should *not* choose one between $t_a R^* t_b$ and $t_{a^*} R^* t_b$ - it should find a way to say that both are represented. Suppose, then, that $v_a R v_b$ represents *both* $t_a R^* t_b$ and $t_{a^*} R^* t_b$. So, $v_a R v_b$ has a composite content, which might be expressed by ($t_a R^* t_b$ & $t_{a^*} R^* t_b$). But clearly such a content is not semantically transparent in the desired way. But the desired semantic transparency seems entailed by **(1)**, and so now it seems that **(1)** is not the case. This conclusion generates a contradiction - in fact, we're trying to determine what would V's content be, supposing that **(1)** obtains in spite of the various (a)-violations it suffers from. And even leaving this problem aside, there would be problems with **(2)** and **(4)**. Suppose that an agent is using the representation V (which includes $v_a R v_b$) to guide their behavior in respect to a T such that $t_a R^* t_b$ is the case but $t_{a^*} R^* t_b$ is not the case. Here, it is legitimate to expect the agent to non accidentally *succeed*: $v_a R v_b$ carries information about $t_a R^* t_b$ which the agent can "use" to appropriately orchestrate their behavior. But if $v_a R v_b$ actually represents ($t_a R^* t_b$ & $t_{a^*} R^* t_b$), then it is false (or extremely non-accurate). The truth value (or degree of accuracy) of V no longer correlates with the agent's behavioral success, and so **(2)** fails to obtain. Given that **(4)** is entailed by **(2)**, **(4)** fails to obtain too. Of course, one could solve *this* specific problem by arguing that the composite content is something that could be best expressed by ($t_a R^* t_b$ *or* $t_{a^*} R^* t_b$). But now the content of V is plainly disjunctive, and falls prey to the disjunction problem in its various forms (cf. Neander 2017) - and notice that, since the original assumption was that systematic (a)- (and (b)-)violations are admissible, the disjunction here seems unrestrained.

Taking stock: that the neurons of neural maps do not map in a neat one-to-one fashion onto stimuli is a serious problem for the defender of NSRs. The absence of the required one-to-one mapping may be enough to claim that neural maps fail to satisfy **(1)**. And, even accepting that the absence of such a map is no reason to deny that **(1)** obtains, there would still be significant problems with **(2)** and **(4)**. It would be at best unclear whether neuronal maps guide their "users" actions in the way structural representations are supposed to carry out their action guiding duties.

Defenders of NSRs might then be tempted to abandon **(2)** and **(4)** to secure the status of cortical maps as NSRVs. In my assessment, this move is technically viable but practically unwise. Recall why NSRs are central in contemporary cognitive neuroscience. They are central because they allow for the happy marriage of mechanical and representational explanations (**§1**). NSRs allow for this marriage because their NSRVs - the causally efficacious bits and pieces that operate within our neurocognitive mechanisms - are imbued with content: their physical shape has important semantic properties in a way such that these semantic properties are allowed to play an active causal role within our neurocognitive systems (**§2**). In the case of NSRVs, then, the semantics itself does the pushing and pulling required by mechanistic explanations. But, assuming that representational accuracy is conducive to pragmatic success, this view *entails* **(2)**: the degree of accuracy between

vehicle and target must be reflected in the agent's odds of pragmatic success. So, abandoning **(2)** means either (i) abandoning the view that representational accuracy is conducive to pragmatic success or (ii) abandoning the view that the content of NSRs plays a causal role compatible with it being a part of mechanistic explanations. Both options are unattractive to the defender of NSRs. Denying (i) is tantamount to admitting that representations are conducive to success regardless of their truth or accuracy value - which is clearly false. But denying (ii) amounts to conceding that the relevant semantic properties of NSRVs do not play any mechanistically relevant causal role - *de facto* undermining the theoretical attractiveness of NSRs for cognitive neuroscience in general and *mechanistic* cognitive neuroscience in particular (cf. O'Brien 2015; Williams and Colling 2017).

In spite of their appearance, then, neural maps are not NSRVs - the structural similarity that they so obviously boast (to our eyes) might not even be really present. And, even if it were present, it would not play the required representational role.

3.3 - Activation spaces

Thus far, I've in an important sense considered only *single* responses, either of individual neurons (**§3.1**) or of multiple neurons topographically organized in neural maps (**§3.2**). Some defenders of NSRs would claim my focus has been too narrow. To see NSRVs one should look at *multiple* responses from a single neuronal structure. For, the relevant (i.e. NSR-underpinning) similarity does not hold between a single activation and a target. Rather, it holds among the structure's entire activation space (i.e. set of all possible responses) and the entire target domain (i.e. the set of all targets the structure is sensitive to). As far as I can see, there are two different arguments for this claim.

The first - and more widespread - variant is ultimately based on the analysis of the behavior of a large class of neurocomputational models (cf. Churchland 1995; O'Brien and Opie 2004; Grush 2004; Shagrir 2012; Williams 2017; Wiese 2016; 2017).[24] Shagrir (2018) usefully expresses the idea common to all these arguments in terms of *input-output modeling*. Let *f* be the function relating the inputs and outputs of a neurocomputational model. In Shagrir's view, such a model is a model of a target domain T if, when v_a and v_b are in the relevant input-output relation specified by *f*, then the corresponding elements in the target domain (t_a and t_b) stand in a relation mathematically described by *f* too. Consider, for example, a model M that takes as input velocities and yields as outputs space traveled in a minute at that velocity. According to Shagrir, M is an input-output model of its target domain just in case it multiplies the input value by 60 - given that $s=vt$ and here $t=60$ seconds. When this happens, the activation space of M - that is, the set of all M's input-output pairings - is clearly structurally similar to the target domain, in a way that seemingly vindicates **(1)**. What, then, about **(2)**-**(4)**? The argument to the effect they obtain vary from account to account - but here I will ignore them, as they won't play any role in my argument below.

The second - and less widespread (to my knowledge, is made only by Williams & Colling 2017) - argument is based on a technique to analyze neuroimaging data known as *representational similarity analysis* (RSA, see Kriegeskorte *et al*. 2008). RSA belongs to the family of "neural decoding" - or, more soberly, multivariate patterns analysis - techniques. These techniques operate

[24] See also (Rutar *et al*. 2022) for a more nuanced - and less structural-representationalist - treatment.

on various types of imaging data to investigate neural representations (e.g. Haxby *et al*. 2001).[25] RSA typically operates on voxels - think of them as three dimensional pixels "making up" the images - and their activation levels. Each activation is treated as a vector of voxels activation levels, so as to compute the distance (i.e. dissimilarity) between each pair of vectors. Based on these distances, the activations are arranged in a *representational dissimilarity matrix*: an activation space expressing the dissimilarity between each pair of activation as a scalar quantity (i.e. a single number; see Kriegeskorte and Kievit 2013 for an accessible introduction to RSA). Importantly, the pattern of similarities and dissimilarities between neuronal responses revealed by the representational dissimilarity matrix "mirrors" the pattern of similarities and dissimilarities expressed by subjects in their similarity judgments (cf. Connolly *et al*. 2012; Ritchie *et al*. 2014; Carlson *et al*. 2014). So, if two responses are similar (i.e. $v_a R v_b$), then their two targets are similar ($t_a R^* t_b$), in a way that seemingly vindicates **(1)**.

Sadly for the defender of NSRs, however, none of these two arguments establishes that **(1)** obtains. Although *both* arguments show a structural similarity holding, they show it holding amongst the *wrong* sorts of things.

Condition **(1)** requires a structural similarity to hold between *a representational vehicle* V *and a represented target* T. But the structural similarities shown above do *not* hold amongst *individual representational vehicles* and *individual targets*. This is especially obvious in the case of the first argument based on input-output modeling. In that case, the structural similarity holds between a *computational process* pairing inputs and outputs and a certain environmental process. But whilst environmental processes can be represented targets, computational processes can't be representational vehicles. Indeed, on a number of accounts of physical computation, computational processes are *defined over* representational vehicles - they're ways in which representational vehicles are manipulated according to certain rules (cf Fodor 1981; O'Brien and Opie 2009; Maley 2021a). This clearly entails that computational processes and representational vehicles are not identical.[26] Thus, the structural similarity revealed by input-output modeling practices can't be invoked to claim that **(1)** obtains.

RSA suffers from a similar problem, though in an attenuated (and less obvious) form. As pointed out by (Davis and Poldrack 2013; see also Coraci 2022 for a philosopher-friendly analysis) it is not entirely clear whether the structural similarity RSA reveals depends on the *representations* involved within a cognitive process or on the *cognitive process* being run during the experimental trial. Two neuronal activation may be similar because they represent similar things - consider, as an example, the neuronal representation of a male face smiling and the neuronal representation of a female face smiling. These two neuronal activations are likely similar because they represent similar things. But now consider the neuronal activation involved in representing a smiling face and a puppy. These neuronal activation might be similar - but, if so, their similarity would not be due to the similarity of their *contents*, but rather to the fact of a same *cognitive process* (say, judging both the smiling face

[25] Pitched at this level of generality, the claim is importantly contested (cf. Ritchie *et al*. 2019; Gessel *et al* 2021). These critical arguments, however, do not apply to RSA, and so I will ignore them here.

[26] *Mutatis mutandis* the same holds for "non-semantic" accounts of computations. In this case, computations are defined over non-representational *digits* or *states*, strings or combinations of which may be representational vehicles when the appropriate conditions are met (cf. Piccinini 2015).

and the puppy good and having a positive affective response to them) operates on them both. These two scenarios can be disentangled with certain appropriate experimental procedures. But the need to disentangle them weakens any inference from the structural similarities shown by RSA and the claim that **(1)** obtains.

Worse still even when the structural similarity revealed through RSA techniques is due to the similarity of the representations (rather than the processes), that structural similarity still fails to support the claim that **(1)** obtains. For the relevant similarity holds between a *representational dissimilarity matrix* and various targets. But representation dissimilarity matrices are not neural vehicles: not only do they abstract away from the spatiotemporal information that is needed to identify vehicles (see Haxby *et al*. 2014, p. 439; Kriegeskorte & Diedrichsen 2019, p. 418) they are in no sense *neural*. They are in no sense "tokened in the head". They're not what vindicating **(1)** requires in this context.

Defenders of NSRs could plausibly object that, whilst I'm correctly pointing out that computational processes and representational dissimilarity matrices are not neural vehicles, they still *reveal something important* about our neural vehicles: they *show* us that our neural representational vehicles being such that certain relevant structural similarities hold between them and their target domain, in a way that vindicates **(1)**.

The objection gets an important point right: computational processes and representational dissimilarity matrices do depict a structural similarity holding amongst certain neural goings-on and certain environmental targets. But even this *depicted* structural similarity fails to satisfy **(1)**. For, that similarity too does not hold amongst *individual* neural vehicles and targets. It holds among the *activation space* of a neural structure (that is, the range of possible responses that neuronal structure can produce) and a range of targets that structure is responsive to. And whilst that activation space does capture something about our brain - namely, how it responds to various stimuli, the activation space itself is not a neural representational vehicle - it *just isn't* something tokened in the brain. It is only a way to compactly *model* what gets tokened in the brain. Or, in the words of cognitive neuroscientist Davies and Poldrack:

> "The dominant theoretical underpinning of representational analyses in most content areas of *f*MRI research is that stimulus *representations can be thought of as points in an n-dimensional space*. This characterization of neural representations in terms of n-dimensional spaces follows from influential work in cognitive psychology on how *psychological representations can often be characterized as points in a representational space* and how a variety of cognitive processes, such as stimulus generalization, categorization, and memory, can be modeled as geometric operations on these representations." (Davies & Poldrack 2013, p. 109, emphasis added).

So, neural representations are *points* in representational spaces - individual responses to individual stimuli, rather than *range* of responses to a stimulus domain.

But can't defenders of NSRs somehow claim that entire activation spaces are representational vehicles, so at allow **(1)** to obtain? No, they cannot. The reason is simple. Activation spaces and representational dissimilarity matrices show us that - for example - if two neuronal responses are similar, then their targets (i.e. what these responses are responses to) are similar too. Rewriting this in the notation used throughout the paper, the result is: $(v_a R v_b) \rightarrow (t_a R^* t_b)$. So, in this case, individual neuronal responses are treated as *vehicle constituents*, and their "targets" are actually target constituents. But individual neuronal responses *just cannot be material constituents* of a larger vehicle V. The reason is simple. As Kirchhoff (2014; 2015) has aptly noticed, constitution is a *synchronic* relation holding between the constituents and the constituted entity.[27] So, if vehicle constituents $v_a...v_n$ constitute vehicle V at time *t*, then $v_a...v_n$ must all be present at *t*. But, in the case at hand, $v_a...v_n$ *cannot* all be present at *t* (and the relevant empirical material *does not* show that they are all present at *t*). The relevant vehicle constituents cannot all be present at *t* because they are various neuronal responses - that is, states - of a *single* neuronal structure or set of structures. And a single structure or set of structures can only token multiple neuronal responses *though time* - it cannot token them all at *t*. Neurons cannot have *multiple* firing rates at the same time. So, multiple neuronal responses cannot be material constituents of a larger vehicle. As a consequence, the structural similarity holding between them and a target domain cannot underpin any NSR.

At this juncture, a defender of NSRs may claim that my arguments overlook the fact that many allow for structural representations to be "made up" by more than a representational vehicle. For example, Shea (2018, p. 118) defines structural representations as: "A collection of representations in which a relation on representational vehicles represents a relation on the entities they represent". So do other defenders of structural representations, including Swoyer (1991) and Ramsey (2007).[28] So, **(1)** need not be narrowly defined in terms of *single* vehicles, as I did in (**§2**). And if so, then the structural similarity shown by activation spaces and representational dissimilarity matrices can satisfy **(1)**.

Defenders of NSRs, however, are *not* free to re-define **(1)** in this way - unless they are willing to withdraw their commitment to mechanistic explanations. For, when **(1)** is so re-defined, the relevant structural similarity holds between *abstracta* - the relevant set of neuronal representations and a relevant set of targets. Now, whilst one could perhaps in principle accept that abstracta function as neural representations (or equivalently, that neural representations really are abstracta), surely abstracta *cannot* be components of mechanisms. Mechanisms and their components are always concrete (cf. Craver 2007). Moreover - and more generally - it is hard not to notice that taking NSRVs to be abstracta seems an entirely *ad hoc* (albeit not uncommon) move: cognitive neuroscientists clearly conceive of representations in very concrete terms; namely as neuronal states (cf. Friston 2005; Mesualm 2008; Villaroja 2017; Backer *et al*. 2022).

---

[27] Importantly, some philosophers are elaborating *diachronic* accounts of constitution (cf. Leuridan & Lodewyckx 2021; Kirchhoff and Kiverstein 2021) that could be used to counter my argument. Sadly, due to space limitations I cannot introduce and comment upon these accounts here. I'm sure the reader will forgive me if I do not further extend this already long paper.

[28] Importantly, however, this point is contested even in the literature on structural representations. Cummins, for example, would never define structural representations in terms of *multiple* representations. This is because, in his view, the parts of a structural representations are *not* representations in their own right (see Cummins 1996, pp. 96-97).

Am I suggesting that the structural similarity - displayed by activation spaces - holding between various neuronal responses and their target domain is representationally idle? Not necessarily. I'm only denying it holds between vehicles and targets so as to underpin NSRs. But it can still have some relevant representational role. For example, it might determine the content of some other type of representation. There are various theories of content based on structural similarity (e.g. Cummins 1996, O'Brien and Opie 2004) - and while these theories often focus on the structural similarity between individual vehicles and targets, nothing prevents us from applying the same idea to *multiple vehicles* and *target domains*.[29] On this view, individual vehicles would get their content in virtue of the structural similarity holding between a set of different vehicles and a target domain. Each vehicle would thus represent what it represents in virtue of its overall role in the similarity. This intuition could be refined in a full-blown theory of content - but doing so is a task for another paper to carry out. But notice that, even if such a theory of content were provided, it *would not* lend support to the claim that activation spaces/neural dissimilarity matrices/multiple neuronal responses are structural representations. There is a clear and obvious difference between a set of vehicles being structurally similar to a set of targets and individual vehicles being structurally similar to individual targets. The former *just isn't what **(1)** requires*.

3.4 - Alternative neural vehicles

Whilst neuronal responses are the *main* neuronal vehicles cognitive neuroscience is interested in, they're not the *only* vehicles cognitive neuroscience is interested in. So, what about *those*? Do they underpin NSRs? No they don't, and for fairly obvious reasons (thankfully!)

Neuronal connections have often been considered representational vehicles. Indeed, connectionists have long argued that connections between neurons may encode information, functioning as our long-term semantic memory (cf. McClelland *et al*. 1986). However, it is commonly accepted that if connections encode information, they do so in a highly distributed way: single connections store multiple "bits" of different contents, and single contents are "spread over" many connections (see Van Gelder 1991; Grush & Mandik 2002). But if this is the case, if really multiple contents are *simultaneously* encoded by *many overlapping connection*, then clearly the mapping from vehicle constituents to target constituents is *many-to-many*; and so **(a)** - and, *a fortiori* **(1)** - fail to obtain for reasons connected with systematic (a)-violations explored in **§3.2** (see also Facchin 2021a for a different argument to the same effect). So, if connections are representational vehicles (which is disputable, see Ramsey 2007), then they're not NSRVs.

Some neuroscientists have recently suggested that global brain states are neural vehicles that represent the agent's overall state (Kaplan & Zimmer 2020; Westlin *et al*. 2023). As far as I can see no one has ever claimed that global brain states are NSRVs. And it is indeed hard to see how they could underpin NSRs: there's clearly no decoupling from an agent's current state! So, global brain states clearly fail to satisfy **(3).**

Lastly, Chemero (2009) and Martinez and Artiga (2021) have argued that *neuronal oscillations* (i.e. patterns of time-locked neuronal activity, see Buzsaki 2006) are representational vehicles. Are they

[29] Indeed, Churchland's (1992) original structural similarity-based account of content was explicitly focused on *multiple* vehicles.

NSRVs? To my knowledge, no one has yet articulated this view. So, I can't provide a detailed analysis of it. However, there are potent *prima facie* reasons to provide a negative answer. Firing patterns instantiated in *different* times can't be constituents of a single vehicle (see **§3.3**), and this seems to prevent many neuronal oscillations from qualifying as NSRVs. Further, the individual neuronal responses "making up" the oscillations would still fail to map on individual targets as seen in (**§3.2**), generating all the problems discussed in that section.

Are there other potential neural vehicles? Not to my knowledge. Sometimes neuroscientists talk about entire neural structures representing (e.g. the fusiform face *area* is sometimes said to represent faces) but it seems clear that it is a metonymic way of speaking: what neuroscientists most plausibly actually mean is that the *responses* or *activations* in various structures represent things. And, there seems to be no other candidate vehicles. Of course, I cannot exclude that new, more sensitive experimental techniques will reveal functionally salient neuronal aggregations *between* the level of the single neurons and that of neuronal maps, or below the level of individual voxels. These may qualify as NSRVs. But surely such vehicles have yet to be identified - so, we can neither observe them right now, nor can they provide a reasonable ground for the "cognitive neuroscience revolution" *right now*.

3.5 - Neural representations unobserved

Time to take stocks, and re-observe the arguments I offered in a less detailed, more holistic, fashion. I have argued that NSRs have not been observed nor manipulated. Indeed, if the arguments I provided thus far are correct, the *bona fide* neural vehicles are not NSRVs.

In (**§3.1**) I focused on individual neuronal responses. I argued that the claim that individual neuronal responses are NSRVs is ambiguous, as it admits three different readings. I have also argued that -regardless of the reading one wishes to adopt - the claim is hardly defensible: individual neuronal response do not seem to break down into interrelated constituents in the desired manner; and indeed the claim that a constituent of an individual neuronal response represents a constituent of the response's target would be a *reductio* of the idea that individual neuronal responses count as NSRVs. Individual neuronal responses, I suggested, are more plausibly interpreted as indicator representations.

In (**§3.1.1**) I focused on a popular argument that could establish that individual neuronal responses are NSRs. The argument allegedly establishes that indicators are a special case of structural representations. So, If individual neuronal responses are indicators, they are (a special case of) NSRs. I attacked that argument on several grounds. Most importantly, I've claimed that individual neuronal responses (and indicators more generally) cannot be (neural) structural representations, as they always fail to satisfy either **(2)** or **(3)**. I have also claimed that such an argument fails to establish that individual neuronal responses (and indicators more generally) are structurally similar to their target in the way requested by **(1)**. And if so, then the vehicles of (neuronal) indicators cannot be (N)SRVs.

In (**§3.2**) I focused on neuronal maps, claiming that they are not NSRVs. First, I have argued that the topological structural similarity holding between neuronal maps and their target domain does not satisfy **(2)**. Contrary to what **(2)** requires (a)- and (b)-violations of that structural similarity do

not decrease an agent's odd of non-accidental success. I also considered other possible structural similarities tying together neuronal maps and their targets, which, not being based on their apparent topological similarity, would be impervious to the argument above. I then ruled this possibility out based on the fact that individual neurons (that is, the relevant vehicle constituents) *do not* map one-to-one onto their targets as required by **(a)**, and so, **(1)** systematically fails to obtain. I considered several ways in which this obstacle could be overcome, and concluded that no one works.

In (**§3.3**) I focus on activation spaces. I claimed that, whilst such spaces actually are structurally similar to their target domains, that structural similarity does not satisfy **(1)**. In fact, **(1)** requires a structural similarity holding between individual vehicles and targets - but activation spaces just cannot be coherently considered being individual vehicles.

Lastly, in (**§3.4**) I considered a number of other alternative neural vehicles that might underpin NSRs, showing that none actually underpins them for fairly obvious reasons.

One final word of clarification about the arguments in (**§§ 3.1 - 3.4**). A neuroscientifically inclined philosopher might be disappointed by the fact that I've discussed a relatively small number of cases. For example, in (**§3.2**) I have never mentioned the primary auditory cortex, even if its map-like, tonotopic, structure is fairly well known. Or, in **§3.1**, I've considered a fairly small number of single cells studies. One might thus worry that I've just discussed too *little* neuroscientific data to support my conclusion that NSRVs have not been observed *by induction*. This would be a fair criticism, were my line of argument supposed to work by induction. But it is not. If correct, my arguments do *not* show that individual neuronal responses, neural maps, and activation spaces all likely fail to satisfy **(1)**-**(4)** because many of them fail to satisfy them. Rather, my arguments show that these entities *cannot* satisfy **(1)**-**(4)**, and thus that they can't *in principle* qualify as NSRVs.

4- Objections and replies

Supporters of the "cognitive neuroscience revolution" will no doubt wish to resist my conclusion. Here I consider some intuitive objections to resist it, showing that they do not really work.

Objection #1: The account of structural representations in (**§2**) is too *demanding*. A less demanding account would reveal that NSRVs are not just present in our brains, but that they have indeed been observed.

Response: Two points in reply. First, we lack an alternative, less demanding, account of NSRs. The account in (**§2**) is widely used (see, for example, Wiese 2016; 2017; Williams 2017; Lee 2019), and the (few) alternative ones are not less demanding - indeed, they're often *more* demanding, as they adopt a *stronger* reading of **(1)** in terms of homomorphisms. Lacking any less demanding alternative, the objection is pretty toothless.

Secondly, it's hard to even imagine the shape of a less demanding alternative. Presumably, the alternative should discard or weaken at least one condition among **(1)** - **(4)**. But my reading of **(1)** is already the weakest one acceptable, and **(1)** cannot be discarded without thereby discarding the very idea of a *structural* representation. My reading of **(3)** is also the weakest reading of decouplability on offer (cf. Chemero 2009, pp. 55-65; Gładziejewski 2015); and **(3)** cannot be

discarded either, as decouplability is an *essential* feature of representations (Haugeland 1991; Orlandi 2020). **(2)** *could* be weakened and discarded - but doing so would hinder the causal relevance of content, in a way that hinders its relevance in mechanistic explanations. Defenders of NSRs can't thus rely on this move - at least, not without abandoning their mechanistic commitments. And since **(2)** entails **(4)**, **(4)** seems off limits too.

Objection #2: NSRs are *action-oriented* representations (Williams 2017; Piccinini 2022).[30] So, they don't represent the world objectively, but in action-salient terms. Hence they are distortive - and thus false or inaccurate - in a way that is *nevertheless conductive to an agent's behavioral success* (cf. Tschantz *et al*. 2020 for a proof of concept). But this clearly runs counter to **(2)**. So, **(2)** should be discarded - and with it, all the arguments above that hinged on **(2)** failing to obtain. So, NSRVs have been observed, after all.

Response: The objection misconstrues the sense in which action-oriented representations are distortive. Sure, they do not represent the world "as is" (whatever this means) - but that's not to say that they represent it falsely or inaccurately. They represent it *through a pragmatic lens*, and what is represented through such a lens can be either accurate/true or inaccurate/false. If I represent a 6 kg stone as throwable, I'm accurately representing the stone in an action oriented manner. If I represent a 666 kg stone as throwable, I'm inaccurately representing it in an action-oriented manner. Compare: if, by looking through red glasses, I see clouds being red, I'm *not misperceiving* - I'm accurately perceiving through red glasses. Thus, the action-oriented nature of NSRs does not force a rejection of **(2)** - or of the "bits" of my arguments based on **(2)** failing to obtain.

Objection #3: The argument in (**§3.2**) is a bit too quick in establishing that individual neurons map onto *many* targets in a way that poses a problem for **(1)**. Neurons need not represent each target to which they respond. Taking a page out of Dretske's (1988) book, one could argue that individual neurons have the *function to represent* only one target, plausibly their preferred one. That might be enough (or at least a substantial step towards) solving the problem with **(1)**, in a way that also avoids the problems with **(2)** neurons mapping onto many targets generated.

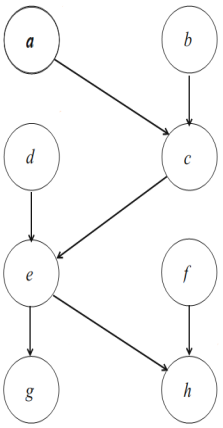
Response: Whilst taking a page out of Dretske's book would solve these problems, the defender of NSRs *can't* rely on Dretske's solution. Dretske assigns functions only after a *learning period*, which stabilizes the function (i.e. determines what the neuron is "supposed to" represent). But real brains have no learning period separate from a non-learning period. Neurocognitive networks are constantly re-organizing and can quickly learn to operate in very odd conditions. As enactivists have repeatedly pointed out, our sensorimotor system can *learn* to operate even in conditions under which sensory and motor signals have been dramatically altered - for exemple, due to one's usage of "inverting goggles" (Hurley 1998; O'Regan 2011). Surely a neuron's learning phase should be over well before the subject is old enough to take part in psychological experiments involving the usage of "inverting goggles"! More generally, it is extremely tricky to assign *well-defined, individual* functions to neuronal areas. Neural functions appear to be multiple, multidimensional, not well-determinate and extremely context dependent (cf. Anderson 2014; Burnston 2016; de Wit &

[30] On the concept of action oriented representations, see (Clark 1997). Curiously, Clark's original example of an action oriented representation is that of Mataric (1991) "spatial map" - a robotic replica of the "spatial map" in the rat's hippocampus. So, it seems that action oriented representations were NSRs all along.

Matheson 2022) - and so will be the contents they ground. Yet, as seen in (**§3.2**), NSRs require reasonably well-determinate contents to function.

Objection #4: My objections to NSRs were hyper-focused on the features of their vehicles. Yet structural representations may not reside at such an "implementational" level of abstraction. They may reside at a higher, "algorithmic" level.[31] Cummins (1989), for example, situated them at the level of program execution. Johnson-Laird (1983) thought of his mental *models* as existing roughly at the same level of abstraction. Similarly, Danks's (2014) suggestion that cognitive representations are graphical *models* sits at a level of abstraction more akin to that of program execution than the implementation level. My arguments are silent about these structural representations and their neural vehicles. So, it fails to rule out NSRs at higher levels of abstraction.

Reply: My reply is simple - Cummins, Danks and Johnson-Laird's models are not *neural* structural representations, so they are unfit to support the "cognitive neuroscience revolution". Moreover, they are not even structural representations in the relevant sense of the term. I maintain that, in the case of structural representations, the vehicle and the target must be structurally similar. And, as far as I can see, this is *not* the case when it comes to the instances of "structural" representations above. Cummins's account is aimed at classical - that is, symbolic - architectures, whose vehicles are *arbitrary* - and in fact, Cummins' (1989) account of structural representations is (roughly) based on the input-output modeling strategy discussed in (**§3.3**). What, then, about Johnson-Laird's mental models and Danks's graphical models? I think that they are (wrongfully) considered as structural representations only because they are presented (to us) in an iconic or iconic-like representational *format*. Consider, for example, the graphical model in **figure 3:**



Caption - Figure 3: A graphical model capturing the statistical dependency relation of some random variables. Made by the author.

Figure 3 represents a simple "Bayesian model" (i.e. a directed acyclic graph), which can be used to model a target phenomenon T. Now, the model - as it is presented to us - surely *seems* a structural representation of T: the nodes *a-h* map one-to-one on aspects of T, and the pattern of arrows "recapitulates" the statistical dependencies in T. But notice that the arrows and nodes we see are

[31] But see Maley (2021b) for an argument to the effect that, in the case of analog representations (including structural ones) the difference between implementational and algorithmic level collapses.

not the vehicle underpinning the model - the vehicle is a complex series of voltages (at the level of the implementation) or "0"s and "1"s (as a higher level) somewhere in my computer. And there is no guarantee (nor any reason to believe) that *it* will be structurally similar to T. Further, the impression of iconicity can be easily dispelled by visualizing the model of **figure 3** in a less graphical (pun intended) format - for example, as the probability distribution p(a, b, c, d, e, f, g, h) = p(g|e) p(h|e, f) p(e|d, c) p(d) p(c|a, b) p(a) p(b).

5 - Conclusions: a dilemma for the cognitive neuroscience revolution.

Suppose my arguments are on the right track: NSRVs have *not* been observed and there is no easy way to avoid this conclusion. This is ill-news for defenders of the "cognitive neuroscience revolution": NSRs are absolutely central to their account (**§1**). So, the question now is: what could revolutionaries do to save their explanatory project? Not much, I fear.

They could try to substitute NSRs with a different type of representation. But this move is unpromising. According to a popular account, there are three basic representational kinds - *icons, symbols and indices* (c.f. Peirce 1931-1958; von Eckart 1996). Now, icons represent by similarity - so neural icons *just are* NSRs, and thus icons are clearly not an option. Symbols represent by stipulation - and so it is not clear if neural symbols can exist: surely no one has *stipulated* the content of our neurons. And even allowing stipulative or stipulation-like processes to take place in the brain (say, as the upshot of a neural signaling game, see Skyrms 2010) the vehicles of neural symbols, being *arbitrary*, do not allow their content to play any causal role within neurocognitive mechanisms. Thus, symbolic representations have no place in mechanistic explanations. Lastly, indices represent in virtue of certain causal relation with their targets - they are indicators. Now, neural indicators surely exist, see (**§3.1**). Yet, it is far from clear they qualify as representations in any robust sense - they seem to function as mere causal mediators in our neurocognitive systems (Ramsey 2003; 2007).

Should then the mechanistic approach to cognitive neuroscience be purged of representational commitments? Some claim this is the case (Kohar *forthcoming*). This, however, would be an *extremely painful* revision of our current neuroscientific practices. Cognitive science is ripe with representational talk, and cognitive neuroscience is no exception. A non-representational mechanistic cognitive neuroscience would thus force us to revise and reinterpret a huge mass of experimental data. It would also force us to find a novel, non-representational lexicon with which to express and communicate the relevant cognitive-scientific findings. This surely is a tall order - one that proponents of the "cognitive neuroscience revolution" do not seem willing to execute.

The only way I see to avoid that non-representational revision, however, seems to be by foregoing one's realistic commitments to NSRs (or at least to NSRVs). The talk of neural maps and models, then, should not be interpreted as referring to real, neurally realized, map- and model- like structures. Rather, neural maps and models are just convenient linguistic tools to understand, track, or make sense of our neurocognitive activities (see Sprevak 2013; Egan 2020; Coelho Mollo 2021 and Cao 2022 for similar views of representations).[32] But to adopt such a construal of NSRs or NSRVs amounts to abandoning one's mechanistic commitments, at least insofar mechanistic

[32] See (Ramsey 2020) for acute criticism of some such accounts.

explanations are *ontic* explanations. But the commitment to mechanism is a core part of the "cognitive neuroscience revolution", and so abandoning it seems to abandon the "cognitive neuroscience revolution" project.

It seems, then, that defenders of NSRs face a dilemma: they either have to let go of their commitment to representationalism to keep their commitment to mechanistic explanations, or *vice versa*. The choice is theirs.

References

Aflalo, T. N., & Graziano, M. S. (2006). Partial tuning of motor cortex neurons to final posture in a free-moving paradigm. Proceedings of the National Academy of Sciences, 103(8), 2909-2914.

Albers, A. M., et al. (2013). Shared representations for working memory and mental imagery in early visual cortex. Current Biology, 23(15), 1427–1431

Anderson, M. L. (2014). After Phrenology. Cambridge, MA.: The MIT Press.

Anderson, M. L., & Champion, H. (2022). Some dilemmas for an account of neural representation: A reply to Poldrack. Synthese, 200(2), 169.

Artiga, M., & Sebastián, M. A. (2018). Informational theories of content and mental representation. Review of Philosophy and Psychology, https:/doi.org/10.1007/s13164-018-0408-1.

Barack, D. L., & Krakauer, J. W. (2021). Two views on the cognitive brain. Nature Reviews Neuroscience, 22(6), 359-371.

Backer, B., et al. (2022). Three aspects of representation in neuroscience. Trends In Cognitive Sciences, 26(11), 942-958.

Bechtel, W. (2008). Mental Mechanisms. Philosophical Perspectives on Cognitive Neuroscience. New York: Routledge.

Bechtel, W. (2014). Investigating neural representations: the tale of place cells. Synthese, 193, 1287-1321.

Bielecka, K., & Miłkowski, M. (2020). Error detection and representational mechanisms. In Smortchkova, J., Dolega, K., & Schicht, T. (eds). What are mental Representations? (pp. 287-317). New York: Oxford University Press.

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. Proceedings of the National Academy of Sciences, 119(3), e2112566119.

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

Born, R. T., & Bradley, D. C. (2005). Structure and function of visual area MT. Annual Review of Neuroscience, 28, 157–189. https://doi.org/10.1146/annurev.neuro.26.041002.131052

Brette, R. (2015). Philosophy of the spike: rate-based vs. spike-based theories of the brain. Frontiers in systems neuroscience, 151.

Brette, R. (2019). Is coding a relevant metaphor for the brain?. Behavioral and Brain Sciences, 42, e215.

Bruineberg, J., & Rietveld, E. (2019). What's inside your head once you've figured out what your head's inside of. Ecological Psychology, 31(3), 198-217.

Buckley, C. L., Kim, C. S, McGregor, S., & Seth, A. K. (2017). The free energy principle for action and perception: a mathematical review. Journal of Mathematical Psychology, 81, 55-79.

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

Buzsaki, G (2006). *Rhythms in the Brain*. New York: Oxford University Press.

Cao, R. (2022). Putting representations to use. Synthese, 200(2), 151.

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. Journal of cognitive neuroscience, 26(1), 132-142.

Chakrabarty, S., & Martin, J. H. (2000). Postnatal development of the motor representation in primary motor cortex. Journal of neurophysiology, 84(5), 2582-2594.

Chemero, A. (2009). *Radical Embodied Cognitive Science*. Cambridge, MA.: The MIT Press.

Churchland, P. M. (1992). *A Neurocomputational Perspective*. Cambridge, MA.: The MIT Press.

Churchland, P. M. (1995). *The Engine of Reason, the Sit of the Soul*. Cambridge, MA.: The MIT Press.

Clark, A. (1997). *Being There*. Cambridge, MA.: The MIT Press

Coelho Mollo, D. (2021). Deflationary realism: Representation and idealisation in cognitive science. Mind & Language, 37(5), 1048-1066.

Connolly, A. C., et al. (2012). The representation of biological classes in the human brain. Journal of Neuroscience, 32(8), 2608-2618.

Coraci, D. (2022). Representations and processes: What role for multivariate methods in cognitive neuroscience?. Rivista internazionale di Filosofia e Psicologia, 13(3), 187-199.

Craver, C. F. (2007). Explaining the brain: Mechanisms and the mosaic unity of neuroscience. Clarendon Press.

Csibra, G. (2008). Action mirroring and action understanding: An alternative account. In P. Haggard, Y. Rossetti, & M. Kawato (Eds.), Sensorimotor foundations of higher cognition (pp. 435–459). Oxford: Oxford University Press.

Cummins, R. (1989). Meaning and mental representation. Cambridge, MA: MIT Press

Cummins, R. (1996). *Representations, Targets, Attitudes*. Cambridge, MA.: The MIT Press.

Danks, D. (2014). Unifying the mind: Cognitive representations as graphical models. Cambridge, MA.: The MIT Press.

Davis, T., & Poldrack, R. A. (2013). Measuring neural representations with fMRI: practices and pitfalls. Annals of the New York Academy of Sciences, 1296(1), 108-134.

Dayan, P., & Abbott, L. F. (2005). Theoretical neuroscience: computational and mathematical modeling of neural systems. MIT press.

De Angelis, G. C., & Newsome, W. T. (1999). Organization of disparity-selective neurons in macaque area MT. Journal of Neuroscience, 19(4), 1398–1415

Dennett, D. C. (1996). *Darwin's Dangerous Idea*. London: Penguin.

de Wit, M. M., & Matheson, H. E. (2022). Context-sensitive computational mechanistic explanation in cognitive neuroscience. Frontiers in Psychology, 4225.

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

Dretske, F. (1988). *Explaining Behavior*. Cambridge, MA.: The MIT Press.

Egan, F. (2020). A deflationary account of mental representations. In J. Smortchkova, K. Dolega, & T. Schlicht (Eds.), What are Mental Representations? (pp. 26–54). Oxford University Press.

Facchin, M. (2021a). Predictive processing and anti-representationalism, Synthese, 199(3-4), 11609-11604.

Facchin, M. (2021b). Structural representations do not meet the job description challenge. Synthese, 199(3), 5479-5508.

Favela, L.; & Machery, E. *forthcoming*. The untenable status quo: the concept of representation in the neural and psychological sciences.

Fodor, J. A. (1981). "The Mind-Body Problem." Scientific American 244 (January 1981). Reprinted in J. Heil, (Ed.) (2004a), Philosophy of Mind: A Guide and Anthology (168–82). Oxford: Oxford University Press

Frisby, S. L., et al. (2023). Decoding semantic representations in mind and brain. Trends in Cognitive Sciences.27(3), 258-281

Friston, K. (2005). A theory of cortical responses. Philosophical Transactions of the Royal Society B: Biological Sciences. 360(1456), 815-836

Gessell, B., Geib, B., & De Brigard, F. (2021). Multivariate pattern analysis and the search for neural representations. Synthese, 199(5-6), 12869-12889.

Gładziejewski, P. (2015). Explaining cognitive phenomena with internal representations: a mechanistic perspective, *Studies in Logic, Grammar and Rhetoric*, 40(1), 63-90.

Gładziejewski, P. (2016). Predictive coding and representationalism, Synthese, 193(2), 559-582.

Gładziejewski, P., & Miłkowski, M. (2017). Structural representations: causally relevant and distinct from detectors. Biology and Philosophy, 32(3), 337-355

Gordon, E. M., et al. (2022). A mind-body interface alternates with effector-specific regions in motor cortex. Nature https://doi.org/10.1038/s41586-023-05964-2

Graziano, M. S. (2011). Cables vs. networks: old and new views on the function of motor cortex. The Journal of physiology, 589(Pt 10), 2439.

Graziano, M. S. (2016). Ethological action maps: a paradigm shift for the motor cortex. Trends in cognitive sciences, 20(2), 121-132.

Graziano, M. S., & Aflalo, T. N. (2007). Mapping behavioral repertoire onto the cortex. Neuron, 56(2), 239-251.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. Behavioral and brain sciences, 27(3), 377-396.

Grush, R., & Mandik, P. (2002). Representational Parts. Phenomenology and the Cognitive Sciences, 1(3), 389-394.

Ha, D., Schmidhuber, J. (2018a). Recurrent world models facilitate policy evolution. In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31 (pp. 2451-2463), Curran Associates.

Ha, D., Schmidhuber, J. (2018b). World models. Preprint. ArXiv:18.0310122.

Hartmann, K., *et al*. (2016). Embedding a panoramic representation of infrared light in the adult rat somatosensory cortex through a sensory neuroprosthesis. Journal of Neuroscience, 36(8), 2406-2424.

Haruno, M., Wolpert, D. M., Kawato, M. (2003). Hierarchical MOSAIC for motor generation. In T. Ono, G. Matsumoto, R. R. Llinas, A. Bethoz, R. Norgren, H. Nishijo, R. Tamura (Eds.), Excepta Medica International Congress System (Vol. 1250), (pp. 575-590). Amsterdam: Elsevier.

Haueis, P. (2018). Beyond cognitive myopia: a patchwork approach to the concept of neural function. Synthese, 195(12), 5373-5402.

Haxby, J.et al. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science, 293(5539), 2425–2430

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. Annual review of neuroscience, 37, 435-456.

Hubel, D., & Wiesel, T. (1968). Receptive fields and functional architecture of monkey striate cortex. The Journal of Physiology, 195(1), 215-243

Hurley, S. (1998). *Consciousness in Action*. Cambridge University Press.

Hutto, D., & Myin, E. (2013). *Radicalizing Enactivism*. Cambridge, MA.: The MIT Press.

Illari, P. (2013). Mechanistic explanation: Integrating the ontic and epistemic. Erkenntnis, 78, 237-255.

Isaac, A. M. (2013). Objective similarity and mental representation. Australasian Journal of Philosophy, 91(4), 683-704.

Itskov, P. M., et al (2011). Hippocampal representation of touch-guided behavior in rats: persistent and independent traces of stimulus and reward location. PLoS ONE 6:e16462. doi: 10.1371/journal.pone.0016462

Johnson-Laird, P. (1983). Mental models. Cambridge, MA: Harvard University Press.

Kaplan, H. S., & Zimmer, M. (2020). Brain-wide representations of ongoing behavior: a universal principle?. Current opinion in neurobiology, 64, 60-69.

Kelso, S. (1995). *Dynamic Patterns*. Cambridge, MA.: The MIT Press.

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. Cognitive Processing, 8(3), 159–166. doi:10.1007/s10339-007-0170-2

Kirchhoff, M. (2014). Extended cognition & constitution: Re-evaluating the constitutive claim of extended cognition. Philosophical Psychology, 27(2), 258-283.

Kirchhoff, M. D. (2015). Extended cognition & the causal-constitutive fallacy: In search for a diachronic and dynamical conception of constitution. Philosophy and phenomenological research, 90(2), 320-360.

Kirchhoff, M. D., & Kiverstein, J. (2021). Diachronic constitution. *Preprint*. http://philsci-archive.pitt.edu/19690/

Kohar, M. (forthcoming). Neural Machines: a defense of non-representationalism in cognitive neuroscience. Springer.

Kohler, E., *et al*. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. Science, 297(5582), 846-848.

Kraus, B. J., Robinson, R. J., White, J. A., Eichenbaum, H., & Hasselmo, M. E. (2013). Hippocampal "time cells": time versus path integration. Neuron, 78(6), 1090-1101.

Kriegeskorte, N. et al. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 4.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. Trends in Cognitive Sciences, 17(8), 401–412.

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the onion of brain representations. Annual review of neuroscience, 42, 407-432.

Kwan, H. C et al. (1978). Spatial organization of precentral cortex in awake primates. II. Motor outputs. Journal of neurophysiology, 41(5), 1120-1131.

Lee, J. (2019). Structural representations and the two problems of content. *Mind & Language*, *34*(5), 606-626.

Lee, J. (2021). Rise of the swamp creatures. Philosophical Psychology, 34(6), 805-828.

Lee, A. Y., *et al*. (2022). The structure of analog representation. Noûs. 2022;1–28. https://doi.org/10.1111/nous.12404

Leuridan, B., & Lodewyckx, T. (2021). Diachronic causal constitutive relations. Synthese, 198, 9035-9065.

Maley, C. (2021a). Analog computation and representation. The British Journal of Philosophy of Science. https://doi.org/10.1086/715031

Maley, C. J. (2021b). The physicality of representation. Synthese, 199(5-6), 14725-14750.

Maley, C. (2023). Icons, magnitudes and their parts. Forthcoming in Critica: Revista Hispanoamericana de Filosofia.

Martin J. H., *et al*. (2000) Impairments in prehension produced by early postnatal sensorimotor cortex activity blockade. J Neurophysiol 83: 895–906.

Martin, J. H., *et al*. (2005). Effect of forelimb use on postnatal development of the forelimb motor representation in primary motor cortex of the cat. Journal of neurophysiology, 93(5), 2822-2831.

Martinez, M. & Artiga, M. (2021). Neural oscillations as representations. The British Journal of Philosophy of Science. https://doi.org/10.1086/714914

Mataric, M. (1991). Navigating with a rat's brain: a neurobiologically inspired model for robot spatial representation. In J. A. Meyer and S. Wilson (Eds.), From Animals to Animats 1 (pp. 169-75). Cambridge, MA.: The MIT Press.

McClelland, J. L., *et al*.. (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Voll. I & II. Cambridge, MA.: The MIT Press.

McLendon, H. J. (1955). Uses of similarity of structure in contemporary philosophy. Mind, 64(253), 79–95

McNamee, D., & Wolpert, D. M. (2019). Internal models in biological control. Annual Review of Control, Robotics, and Autonomous Systems 2, 339-364

Mesulam, M. (2008). Representation, inference, and transcendent encoding in neurocognitive networks of the human brain. Annals of neurology, 64(4), 367-378.

Morgan, A. (2014). Representations gone mental. Synthese, 191(2), 213-244.

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

Moser E. I., Kropff E., & Moser M. B. (2008). Place cells, grid cells, and the brain spatiotemporal representation system, *Annu. Rev. Neuroscience*, 31, 69-89, doi: 10.1146/annurev.neuro.31061307.090723

Neander, K. (2017). A Mark of the Mental. Cambridge, MA.: The MIT Press.

Nieder, A., Diester, I., & Tudusciuc, O. (2006). Temporal and spatial enumeration processes in the primate parietal cortex. Science, 313(5792), 1431–1435.

Nirshberg, G., & Shapiro, L. (2020). Structural and Indicator representations: a difference in degree, not in kind. Synthese, https://doi.org/10.1007/s11229-020-02537-y

O'Brien, G. (2015). How does mind matter? Solving the content causation problem. In T. K. Metzinger & J. M. Windt (Eds.), Open mind. Frankfurt am Main: MIND Group. doi:10.15502/9783958570146

O'Brien, G., & Opie, J. (2009). The role of representation in computation. Cognitive processing, 10, 53-62.

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

O'Regan, K. (2011). *Why doesn't red sound like a bell*. New York: Oxford University Press.

Orlandi, N. (2020). Representing as coordinating with absence. In J. Smortchkova, K. Dołega, T. Schlicht (Eds.), What Are Mental Representations?, (pp. 101-134). New York: Oxford University Press.

Peirce, C. S. (1931–1958). Collected papers of Charles Sanders Peirce. In: P. Hartshorne, P. Weiss, & A. Burks (Eds.) (Vols. 1–8). Cambridge, MA: Harvard University Press

Penfield, W., & Boldrey, E. (1937). Somatic motor and sensory representation in the cerebral cortex of man as studied by electrical stimulation. Brain, 60(4), 389-443.

Penfield, W., & Rasmussen, T. (1950). The cerebral cortex of man; a clinical study of localization of function. New York: Macmillan

Pezzulo, G. (2008). Coordinating with the future: the anticipatory nature of representation. Minds and Machines, 18, 179-225.

Piccinini, G. (2015). *Physical Computation*. New York: Oxford University Press.

Piccinini, G. (2020a). *Neurocognitive Mechanisms*. New York: Oxford University press

Piccinini, G. (2020b). Nonnatural mental representations. In G. Smortchkova, K. Dolega, & T. Schlicht (Eds.), What Are Mental Representations? Oxford University Press.

Piccinini, G. (2022). Situated neural representations: solving the problems of content. Frontiers in Neurorobotics, 16, http://doi.org/10.3389/fnbot.2022.846979

Pickering, M. J., & Clark, A. (2014). Getting ahead: forward models and their place in cognitive architecture. Trends in cognitive sciences, 18(9), 451-456.

Poldrack, R. (2020) The physics of representation. *Synthese*, *199*, 1307-1325.

Quiroga, R. et al. (2005). Invariant visual representation by single neurons in the human brain. Nature, 435(7045), 1102-1107.

Ramsey, W. (2003). Are receptors representations?, Journal of Experimental & Theoretical Artificial Intelligence, 15(2)

Ramsey, W. (2007). *Representation Reconsidered*. Cambridge: Cambridge University Press.

Ramsey, W. (2016). Untangling two questions about mental representation. New Ideas in Psychology, 40, 3-12.

Ramsey, W. (2020). defending representation realism. In J. Smortchkova, K. Dolega, & T. Schlicht (Eds.), What are Mental Representations? (pp. 54-84). Oxford University Press.

Ritchie, J. B., Tovar, D. A., & Carlson, T. A. (2015). Emerging object representations in the visual system predict reaction times for categorization. PLOS Computational Biology, 11(6), e1004316.

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. The British journal for the philosophy of science.

Rizzolatti, G., & Sinigaglia, C. (2023). *Mirroring Brains*. New York: Oxford University Press.

Rosche, W., & Sober, E. (2019). Disjunction and distality: the hard problem for purely probabilistic causal theories of mental content. Synthese, https://doi.org/10.1007/s11229-019-02516-y.

Rutar, D., Wiese, W., & Kwisthout, J. (2022). From representations in predictive processing to degrees of representational features. Minds and Machines, 32(3), 461-484.

Schieber, M. H. (2001). Constraints on somatotopic organization in the primary motor cortex. Journal of neurophysiology, 86(5), 2125-2143.

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

Seth, A. K. (2015). The cybernetic bayesian brain. In T. Metzinger, J. Windt (eds.). Open MIND, 35. Frankfurt am Main, The MIND Group. https://doi.org/10.15502/9783958570108

Shagrir, O. (2012). Structural representations and the brain. The British Journal for the Philosophy of Science. 63(3), 519-545

Shagrir, O. (2018). The brain as an input–output model of the world. Minds and Machines, 28, 53-75.

Shea, N. (2018). Representation in Cognitive Science. New York: Oxford University Press.

Silberstein, M., & Chemero, A. (2013). Constraints on localization and decomposition as explanatory strategies in the biological sciences. Philosophy of Science, 80(5), 958-970.

Silberstein, M. (2021). Constraints on localization and decomposition as explanatory strategies in the biological sciences 2.0. In M. Viola, F. Calzavarini (Eds.) Neural Mechanisms: new challenges in the philosophy of neuroscience. Springer.

Skyrms, B. (2010). Signals: Evolution, learning, and information. OUP Oxford.

Sprevak, M. (2013). Fictionalism about neural representations. The Monist, 96(4), 539-560.

Sterling, P., & Laughlin, S. (2015). Principles of neural design. MIT press.

Sun, C., Yang, W., Martin, J., & Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. Nature neuroscience, 23(5), 651-663.

Swoyer, C. (1991). Structural representation and surrogative reasoning. Synthese. 87(3), 449-508

Tani, J. (2007). On the interactions between top-down anticipation and bottom-up regression. Frontiers in Neurorobotics, 1:2.

Tani, J. (2016). *Exploring robotic minds*. New York: Oxford University Press.

Thomson, E., & Piccinini, G. (2018). Neural representations observed. Minds and Machines, 28, 191-235.

Tootell, R. B., Switkes, E., Silverman, M. S., & Hamilton, S. L. (1988). Functional anatomy of macaque striate cortex. II. Retinotopic organization. Journal of neuroscience, 8(5), 1531-1568.

Tschantz, A., Seth, A. K., & Buckley, C. L. (2020). Learning action-oriented models through active inference. PLoS computational biology, 16(4), e1007805.

Van der Weel, F. R., Sokolovskis, I., Raja, V., & van der Meer, A. L. (2022). Neural Aspects of Prospective Control through Resonating Taus in an Interceptive Timing Task. Brain Sciences, 12(12), 1737.

Van Gelder, T. (1991). What is the “D” in “PDP”? A survey of the concept of distribution. In W. Ramsey, S. P. Stich, D. E. Rumelhart (Eds.). Philosophy and Connectionist Theory, New York: Routledge.

Vilarroya, O. (2017). Neural representation. A survey-based analysis of the notion. Frontiers in psychology, 8, 1458.

Von Eckardt, B. (1996). What is Cognitive Science?. Cambridge, MA.: The MIT Press

Wassermann, E. M., et al. (1992). Noninvasive mapping of muscle representations in human motor cortex. Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section, 85(1), 1-8.

Westlin, C.,et al. (2023). Improving the study of brain-behavior relationships by revisiting basic assumptions. Trends in Cognitive Sciences. 27(3), 246-257.

Wiese, W. (2016). What are the contents of representations in predictive processing?. Phenomenology and the Cognitive Sciences, 16, 715-736.

Wiese, W. (2017). *Experienced Wholeness*. Cambridge, MA.: The MIT Press.

Williams, D. (2017). Predictive processing and the representation wars. Minds and Machines, 28(1), 141-172.

Williams D., & Colling L. (2017). From symbols to icons: the return of resemblance in the cognitive science revolution, Synthese, 195(5), 1941-1967

Wood, E. R., et al. (1999). The global record of memory in hippocampal neuronal activity. Nature, 397(6720), 613-616.

Wood, E. R., et al. (2000). Hippocampal neurons encode information about different types of memory episodes occurring in the same location. Neuron, 27(3), 623-633.

Woolsey et al. (1952). Patterns of localization in precentral and" supplementary" motor areas and their relation to the concept of a premotor area. Research publications-Association for Research in Nervous and Mental Disease, 30, 238-264.

Yarrow, S., *et al*. (2014). Detecting and quantifying topography in neural maps. PloS one, 9(2), e87178.