--PREPRINT--

Neural Representations Without Original Content: The Case from Content Multiplicity

Ori Hacohen & Gal Vishne

Abstract

Current neuroscientific explanations regularly refer to internal “neural representations” in explaining cognitive phenomena, yet the nature of these representations remains obscure. Many maintain that these are natural representational entities which carry "original", or "intrinsic" contents. We argue against this naturalistic view and aim to show that neural representations are partially dependent on subjective explanatory considerations. First, we show that neuroscientists routinely regard the same neural state as a representation of multiple distinct contents. We maintain that such content multiplicity is a characteristic feature of the increasingly prevalent population approach in neuroscience. Second, we argue that naturalistic theories are inadequate to explain content multiplicity. This follows from a crucial property of any representation, namely, that it is an exclusive relation. Naturalistic theories are incapable of defining one exclusive content from a multitude of options without turning to subjective considerations. Therefore, the only way to account for content multiplicity is to accept that subjective considerations have a constitutive role in defining the contents of neural representations.

Keywords: Neural representation; Neural coding; Philosophy of neuroscience; Philosophy of mind; Theories of content

1 Introduction

Contemporary neuroscientific explanations regularly refer to internal neural entities as “representations” of distal content. Talk of “information processing” or “coding” in the brain is ubiquitous throughout neuroscientific practice, and it is generally assumed that such processes require internal "neural representations". The nature of these representations, though, has been the focus of substantial debate in both the neuroscientific (e.g., Brette, 2019; Baker et al., 2022; Elber-Dorozko & Loewenstein, 2023) and philosophical communities (e.g., Egan, 2014; Neander, 2017; Shea, 2018; Piccinini, 2022; Hacohen, 2022). Arguably, the most fundamental aspect of this debate is the

question of whether or not such contents are somehow dependent on the subjective considerations of researchers.

Any physical representation is essentially a vehicle which carries content or information. Such representational vehicles are common in our everyday lives. A fuel gauge, for example, carries information about the amount of fuel in the car's tank, and a stop sign carries the directive content "stop". The content of these everyday representations is always dependent upon, and at least partially defined, by *us* – the cognitive agents which create and use them.[1] On the other hand, when turning to mental representations and thoughts, it is commonly assumed that their contents are *not* dependent on the intentions or conventions of cognitive agents. Instead, mental representations are presumed to have "original content", that is, content which is *intrinsic* to the representational vehicle, and entirely independent of external subjective considerations. Explicating what this means and how original content comes about is one of the foundational questions in the philosophy of mind, and philosophers have offered a large variety of *naturalistic* theories of content, which aim to account for original contents in non-intentional, non-semantic terms (Millikan, 1984, 1989; Dretske, 1988, 1995; Fodor, 1987, 1990).

But what about *neural representations*, ostensively defined as the representations which contemporary neuroscientific explanations allude to? Are these also natural representations, with original content (in which case they may help in explaining natural mental content), or do they rely on subjective considerations, akin to other more conventional representations? Here too, the mainstream approach has largely followed the first route (e.g., Neander, 2017; Shea, 2018; Piccinini, 2022). Still, it has also been argued that the representations which neuroscientists allude to are at least partially defined by the explanatory context in which they are posited, and the subjective considerations of researchers, and as such do not carry original contents (Egan, 2014, 2018; Cao, 2022; Hacohen, 2022). Such accounts are usually referred to as *pragmatic* theories of representation.

Practically all theories of representation, naturalistic and pragmatic alike, agree that for a neural entity v to be a representation of some feature of the world X, there must be some objective correspondence between states of v and states of X (also referred to as correlation, or *tracking*). Yet, it is also agreed that the existence of such a correspondence is not sufficient to define representation. Naturalistic theories, arguing for original

[1] With regards to the fuel gauge's dependence on us, see (Dretske, 1988, pp. 59-60).

content, propose additional naturalistic conditions, such as: appealing to the teleological function of the neural vehicle (Neander, 2017), its effect on downstream components (Millikan, 1989), its situated, embedded nature (Piccinini, 2022), or its structural similarities with the represented content (Gallistel, 1990). Pragmatic theories, on the other hand, maintain that these conditions remain insufficient, and argue that the subjective explanatory context is indispensable to defining neural representations. In this paper we offer an argument in favor of this latter view.

Our argument appeals to a feature of neural representations which we refer to as "content multiplicity", whereby the same exact representational vehicle carries multiple distinct contents. The famous "rabbit-duck" sketch[2] depicted in Figure 1 can be regarded as an intuitive example of content multiplicity. In section 2, we show that in contemporary neuroscientific practice, content multiplicity has become a characteristic feature of neural representations. Then, in section 3, we explain why content multiplicity is inconsistent with a naturalistic approach to neural representation, and how it necessitates incorporating subjective considerations into determining each of the multiple different contents. Section 4 presents the conclusions following from this argument – namely, neural representations do not have original contents and are instead partially dependent on the subjective considerations of researchers.[3]

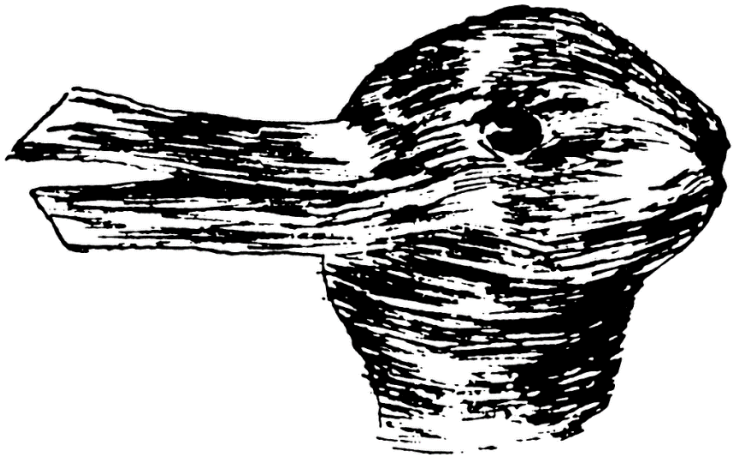

Figure 1. The rabbit-duck sketch. An example of a single physical vehicle which can be regarded as carrying both the content RABBIT and the content DUCK. For example, consider the elongated shapes at the left side of

[2] The rabbit-duck sketch was made famous mostly thanks to Wittgenstein (1953).
[3] Importantly, our argument and conclusions are limited to neural representations, as used by contemporary neuroscientists. We do not aim to show there is no original content, or that mental representations do not carry original content.

the image: when interpreting the image as a rabbit, these correspond to the ears, and when interpreting it as a duck, these correspond to the beak.

2 Content Multiplicity in Contemporary Neuroscience

2.1 The Population Doctrine

For over a century now the single neuron has been treated as the fundamental unit of computation and representation. This approach, referred to as the "neuron doctrine", is steadily being replaced by a new "population doctrine", which posits that the basic representational unit, and the basic unit of computation is not a single neuron, but groups of neurons (Yuste, 2015; Saxena & Cunningham, 2019; Ebitz & Hayden, 2021). These ideas have been developed for many years in the theoretical literature (e.g., Averbeck et al., 2006). However, it is only in recent years that this shift has been ushered into the forefront of neuroscientific research. This change is attributed to two factors: first, the invention of high throughput recording techniques, which enable recording of hundreds, or even thousands of neurons simultaneously, and second, analytical advancements which enable researchers to make sense of this unprecedented amount of data (Cunningham & Yu, 2014). As Saxena and Cunningham (2019, pp. 103-104) state: "These scientific findings are, first, reshaping the way the field thinks about computation, and, second, fundamentally population-based. Taken together, these two features point to a future where the central scientific theme is not the neuron doctrine, but the neural population doctrine".[4]

The core notion behind the population doctrine is the neural state-space (Figure 2): An abstract space where each axis marks the activity of a single neuron, and thus the population activity at a specific moment in time corresponds to a point in this space (Figure 2a). As neural activity is inherently dynamic, it is customary to speak of trajectories within this space (Figure 2b).

With regard to neural representation, the population doctrine is in agreement with the traditional neuron doctrine about requiring a correspondence between the neural state and the world state. However, the two doctrines differ in their approach to this required *tracking relation*. Under the neuron doctrine, neuron n is said to track world feature X if the firing rate of the neuron is systematically related to the value of X. For example (Figure 2c), X can be the color of a specific patch in the visual field – if the firing

[4] For more reviews on the centrality of population representations to contemporary neuroscience see also: Fusi et al., 2016; Vyas et al., 2020; Barack & Krakauer, 2021.

rate of n is higher when the patch is orange (stimulus 1) compared to when it is blue (stimulus 2), we will say the neuron is tracking color (or that it is sensitive to color). Tracking by a neural population builds on this definition and extends it to the neural state-space (Figure 2d) – population p is said to track feature X if the population position along some axis within the state-space covaries systematically with the value of X.[5] This axis correspondence is called a "coding dimension" (Ebitz & Hayden, 2021). Put differently, the tracking relation relevant to population coding is captured by *projecting the neural activity* to the coding dimension, i.e., focusing only on the neural activity parallel to the coding dimension and ignoring all other dimensions.

From this it follows directly that a population can *simultaneously track multiple distinct features*, by using orthogonal coding dimensions, each corresponding to a distinct feature (Figure 2e).[6] Thus, tracking multiple distinct features is an inherent property of the population view. Importantly, coding dimensions typically rely on the activity of many neurons, neither of which individually correlate with X. This means that the relevant physical vehicle necessarily includes the population as a whole and is not reducible to isolated single neurons. To give a schematic example, assume we have two neurons, n_1 and n_2, both firing in relation to two distinct world features X and Y, such that:

$$n_1 = \frac{1}{2}X + \frac{1}{2}Y$$

$$n_2 = \frac{1}{2}X - \frac{1}{2}Y$$

Neither n_1 nor n_2 tracks X or Y in isolation since it is impossible to know what the true value of X or Y based on each neuron's firing alone. However, examining both neurons together, as a population (n_1, n_2), we can extract precisely the values of X and of Y using the orthogonal coding dimensions (1, 1) and (1, -1), since:

$$X = n_1 + n_2$$

$$Y = n_1 - n_2$$

That is, the neural population tracks two distinct contents, X and Y, using the same set of states, defined by the firing rate of the population (n_1, n_2).

[5] Coding can also be done by more than one dimension, but for simplicity we focus on a single coding dimension which is very common in the literature.

[6] Orthogonality is important because it means that the correspondence between the population and one feature is *entirely independent* of its correspondence with other features. This allows us to account for cases where the population misrepresents one feature but still correctly represents other features (see section 3 for additional discussion of misrepresentation).

To conclude, we have shown that the population doctrine allows for *simultaneous tracking* of multiple distinct features. As mentioned in section 1, there is a consensus in the philosophical community that tracking, on its own, does not amount to representation. Yet, specifying the relevant tracking relation(s) *is* enough to explicate the relevant physical vehicle, and the content(s) being tracked. Thus, the multiplicity of tracking inherent to the population approach enables the *representation of multiple distinct contents by a single physical vehicle*. As we show in the following section, such cases of content multiplicity are central to contemporary neuroscience.

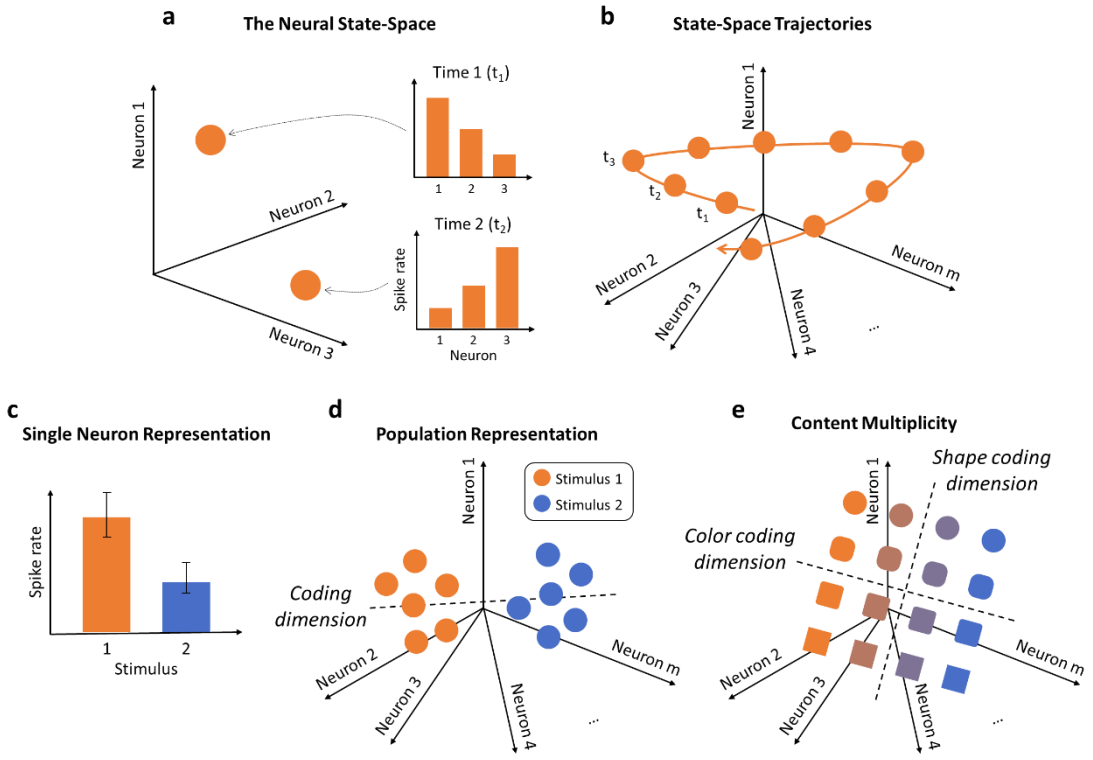

Figure 2. The population doctrine. (**a**) Illustration of the neural state-space: Each axis corresponds to the activity of a single neuron. The simultaneous activity of multiple neurons can either be represented as a point in this space, or as a histogram across neurons (insets on the right). (**b**) State-space trajectories depict the temporally evolving state of the population, e.g., from the onset of a stimulus until it is terminated. (**c**) Single neuron representation: The bar plot depicts the theoretical response of a visual neuron responding to an orange or a blue stimulus. Since the neuron responds significantly more to the orange stimulus, we would say that the neuron is color sensitive. (**d**) Population representation: Theoretical visual population response (encompassing many neurons), under the same experimental scenario as c. Each point's location corresponds to the population activity in response to a single presentation of the orange or blue stimuli (with the color corresponding to the presented stimulus). The "clouds" formed by each of the stimuli are separable from each other in state-space. The direction of this separation is the color coding dimension (marked by a dashed line). (**e**) Content multiplicity: Extending the example from c-d, the subject is now presented with stimuli varying in both color (from orange to blue) *and* shape (from circle to square). Color information and shape information are tracked by distinct and orthogonal coding dimensions.

2.2 Content Multiplicity Example: Mante et al., 2013

To exemplify the centrality of content multiplicity to contemporary practice, we turn to a highly influential study by Mante et al. (2013), examining context-dependent decision making. The task in this study is based on the random dot kinematogram (RDK), which has provided invaluable insight to our understanding of decision making and other related processes in multiple species (e.g., Shadlen & Newsome, 1996; Gold & Shadlen, 2007; Hanks & Summerfield, 2017). In the RDK task subjects view an array of flashing dots, constructed so that in each frame a certain percentage of the dots are shifted slightly to the same direction (usually left\right), so that they generate the perception of coherent movement, while the other dots are reallocated to random spots within the array. Subjects are then required to indicate the direction of coherent movement. Task difficulty is controlled by varying the percentage of dots moving to the same direction, referred to as "motion coherence".

The novelty of this study was introducing another dimension to the task: color (Figure 3, reproduced from Mante et al., 2013). In the standard RDK task all dots are presented in the same color, yet in the Mante et al. (2013) task, dots were either red or green, with different proportions in each trial (this is referred to as "color coherence"). In each trial, subjects (two macaque monkeys) were required to integrate and perform a decision based on just one of the input streams (either motion or color), depending on the trial context, cued in the beginning of each trial and continuously visible throughout (Figure 3a). Both motion and color coherence varied on a trial-by-trial basis (Figure 3b). While monkeys were largely successful in responding based on the cued feature (Figure 3c,f), the un-cued feature still influenced behavioral responses (Figure 3d,e). Thus, there is evidence that both motion and color affected behavior in both types of trials (motion context or color context).

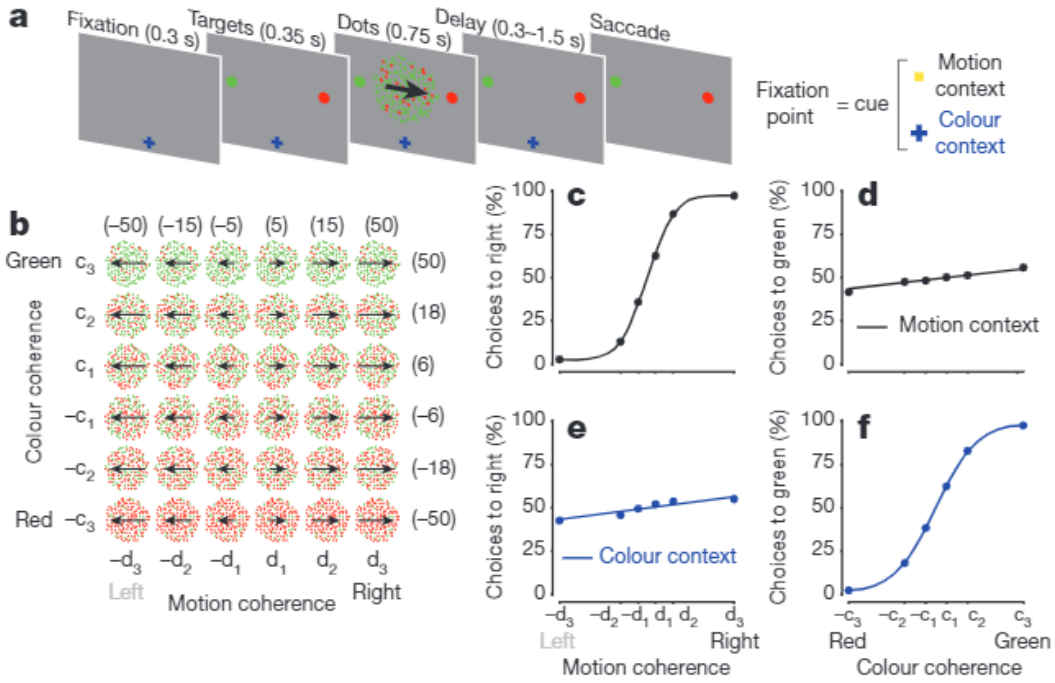

Figure 3. Task structure and behavioral performance (reproduced from Mante et al., 2013). (**a**) Trial structure: monkeys were tasked with discriminating the motion direction or the color of a noisy input stream, consisting of a series of arrays of flashing dots (RDK, see main text). Each trial began with the presentation of a context cue, a colored fixation cross indicating the relevant feature in this trial (motion\color), followed by two target points, used for reporting the decision later in the trial. Then, the main stimulus was presented for 750 ms, and after a variable delay, the fixation cross disappeared, serving as the prompt for the monkeys to report the dominant motion\color, by moving their eyes to the appropriate target. (**b**) Stimuli details: each RDK was characterized by different degrees of "motion coherence" and "color coherence", randomized on each trial. Motion coherence indicates the percentage of points moving coherently, with all other dots moving randomly (positive – rightwards motion, negative – leftwards motion; values used in the study shown on the top row). Color coherence indicates the percentage of dots with the same color (adjusted so that 0% corresponds to an equal number of dots from both colors and ±100% corresponds to arrays with all dots in the same color; values used in the study shown on the right column). (**c-f**) Behavioral performance (monkey A): (**c-d**) Motion context influence on responses of: (**c**) motion information (cued feature) (**d**) color information (un-cued feature) (**e-f**) Color context influence on responses of: (**e**) motion information (un-cued feature) (**f**) color information (cued feature). In both contexts, the cued feature (c, f) showed a clear influence on responses, taking the form of sigmoid function, typical of such psychophysical tasks. The task of the monkeys was to ignore the other, un-cued, feature, yet it still made a small impact. This can be seen by the small slope of the lines: in the motion context high green\red coherence led to slightly higher green\red responses (d), and similarly for high left\right coherence in the color context (e).

Mante et al. (2013) recorded from prefrontal cortex while monkeys were performing this task. They found single neurons sensitive to motion, color, context, and the choice of the animal, yet crucially – most neurons showed *mixed selectivity*, with two or more of these variables modulating their response.[7]

Despite this deep entanglement at the single neuron level, representation at the level of the population was highly separable, with roughly orthogonal coding dimensions for color, motion, and choice information (Figure 4, reproduced from Mante et al., 2013): Each row of Figure 4 focuses on trials from one of the contexts (motion\color discrimination), presenting *the population response trajectories from the same trials* in all three panels. Panels within each row look different because in each of them the trials are grouped according to *different trial features* or viewed from *different directions in state-space*.[8] In all panels, trajectories start from the same point before the onset of the RDK (purple dot), then, after the onset, the trajectories begin to diverge according to the trial specific characteristics. In both contexts, the leftmost panel (Figure 4a,d) shows that the neural population codes the trial's MOTION coherence, and the rightmost panel (Figure 4c,f) shows that it codes the trial's COLOR coherence. As these are the same trials, this shows that motion and color information become available at the same time in the same neural population. Thus, this highly influential study reveals *a single neural state which concurrently represents multiple distinct contents*.

To expand on how coding of motion and color is depicted in these panels, let us first examine the motion context trials (top row, Figure 4a-c): Figure 4a depicts the population response grouped by the trial *motion coherence* and the animal's choice. The responses are projected to two axes denoted as choice (x-axis) and motion (y-axis), which is a shorthand for marking that these are the coding dimensions for choice and motion. To see how the y-axis is the MOTION coding dimension, focus only on the part of the neural trajectory that is parallel to the y-axis – as the trial progresses, leftwards and rightwards motion trials diverge to opposing directions, and the magnitude of this divergence is proportional to the motion coherence (darker lines, corresponding to stronger coherence, show larger deflections from the choice axis). That is, projecting responses to the y-axis shows responses perfectly ordered according to the motion coherence, indicating

[7] See Fusi et al., 2016 for a more general discussion of mixed selectivity in single neurons and how this relates to multivariate representation.

[8] The full population response is highly multidimensional – more than 1,000 neurons were recorded in the study, and this is only a small part of the response of the region. Therefore, for visualization reasons, the population trajectory must be projected to only two axes at a time. The axes used for this purpose determine the state-space viewing direction.

systematic covariance of motion information and the population position along the y-axis, the hallmark of population coding (see section 2.1).[9,10] Figure 4b shows the same trials, but from a *different direction* – the x-axis continues to code for choice, but the new y-axis is now the COLOR coding dimension. Yet, it is impossible to see why this is the case at this point, since the trials are still grouped by motion coherence. Finally, Figure 4c takes these same trials, from the same direction as Figure 4b, and re-groups them according to their *color coherence*. This regrouping reveals a similar pattern to the motion coding observed in Figure 4a, only now with respect to color: Projecting the population responses to the y-axis reveals responses perfectly ordered with respect to the color coherence, indicating that the y-axis in Figure 4b-c is indeed the COLOR coding dimension. Turning to the color context (bottom row, Figure 4d-f) shows a similar pattern, with Figure 4d showing MOTION coding and Figure 4f showing COLOR coding, achieved *simultaneously by the same neural population*.[11] Thus, these results indicate that that the prefrontal cortex population, via the exact same set of states (defined by the firing rates of the entire population), amounts to a single representational vehicle carrying two distinct contents: MOTION and COLOR.

[9] Similarly, to see why the x-axis is the CHOICE coding dimension, focus only on the part of the trajectory that is parallel to the x-axis – once the trial begins, trials where the choice was 1 (leftward motion) diverge to the left and trials where the choice was 2 (rightward motion) diverge to the right. Thus, from an early point in the trial, the position of the neural response along the x-axis is indicative of the animal's choice.

[10] In this figure motion\color information seems to increase until a certain point, when the trajectory inverts direction and begins ascending towards the pre-trial uninformative state. A more recent analysis of this data (Aoi et al., 2020) showed this is an artifact of focusing on a single coding dimension, and information about both aspects remains accessible in other dimensions, where the trajectory exhibits rotational dynamics.

[11] Movement along the x-axis (choice axis) differs between contexts, since the choice is based on motion information in the top row, and color information in the bottom row.

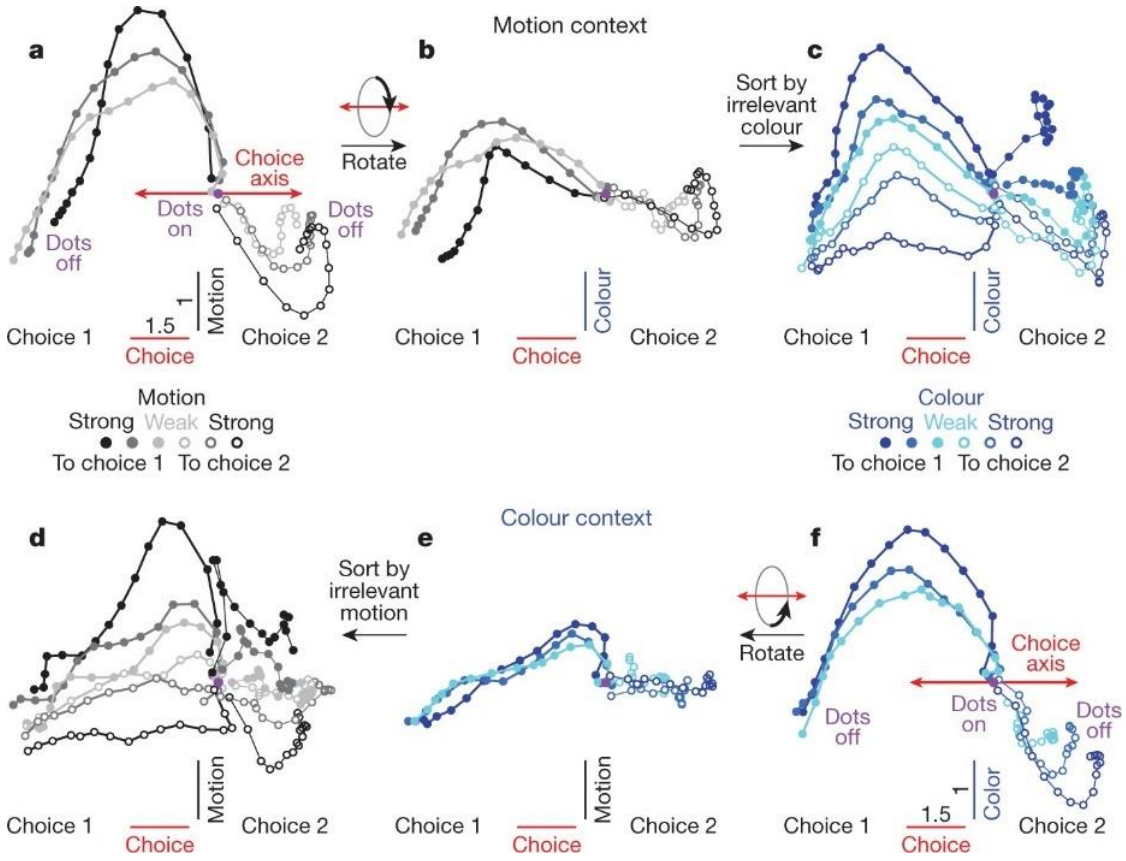

Figure 4. Neural population state-space trajectories reveal coding of multiple task-related contents by the same neural state (reproduced from Mante et al., 2013). All panels depict the mean neural population state-space trajectories (from monkey A), constructed by averaging only correct trials, grouped by the animal's choice (empty\filled dots) and the trial motion coherence (marked by gray\black shades; a, b, d) or color coherence (blue\cyan shades; c, e, f). Stronger coherence is indicated by darker colors. Since the full population response is highly multidimensional (see footnote 8), the figure depicts the population responses projected to two axes (units shown in panels a, f) – the x-axis shows coding for the animal's choice (choice coding dimension) and the y-axis is either the motion coding dimension (a, d, e) or the color coding dimension (b, c, f). Trajectories are shown from 100 ms after the onset of the RDKs (marked by a purple dot), to 100 ms after the offset. Dots on the trajectory mark the population state-space response in intervals of 50 ms. Each row depicts responses to *the same trials*, viewed from different directions or grouped according to different trial characteristics. Top row – motion context: (**a**) grouping by choice and motion (cued feature), projected to the choice and motion axes. (**b**) grouping by choice and motion, projected to the choice and color axes (rotated version of a). (**c**) grouping by choice and color (regrouping to the un-cued feature), projected to the choice and color axes. Bottom row – color context: (**d**) grouping by choice and motion (un-cued feature), projected to the choice and motion axes. (**e**) grouping by choice and color (regrouped to the cued feature), projected to the choice and motion axes. (**f**) grouping by choice and color, projected to the choice and color axes (rotated version of e).

3 Why Content Multiplicity Implies Non-Original Contents

3.1 Representation as an Exclusive Relation

To understand why content multiplicity is incompatible with a naturalistic view of representation we must first turn our attention to a fundamental aspect of representation in general – that it is an *exclusive* relation. We will say that a relation R is an *exclusive* relation between v and X if v being in relation R with X means that v is in relation R with *only* X. To give a simple example, a "legal marriage" relation is exclusive (in most countries). If v is "legally married to" X then, as such, v is necessarily *not* "legally married to" Y (assuming $X \neq Y$). To be "legally married to" X is to be "legally married to" *only* X. A "friendship" relation, on the other hand, is *not* exclusive. If v is "a friend to" X then, as such, v can also be "a friend to" Y. To be "a friend to" X is *not* to be "a friend to" *only* X. And, at least with respect to exclusivity, "representation" is more akin to "legal marriage" than it is to "friendship". Representation, or "content-bearing", is an exclusive relation. To represent X is to represent *only* X.

One way to understand this point is through the concept of *intentionality*, commonly regarded as *the* defining property of representation.[12] Intentionality is often described as "aboutness" – the property of one thing being *about* something else. The introduction of this term into modern philosophy is attributed to Brentano's (1874) book "Psychology from an Empirical Standpoint". There, Brentano characterized intentionality as "direction towards an object". By adopting this characterization, we can get an intuitive sense for the exclusivity of representation. For, if X and Y are different, then v's "direction towards" X is, as such, necessarily *not* a "direction towards" Y. A "direction towards" X is a "direction towards" *only* X, i.e., "direction towards" is an exclusive relation. Now, if this is how we think of representation, then it is an exclusive relation just as well.

Another way to appreciate the exclusivity of representation is through the veridicality of content-bearing states. A physical vehicle v being a representation of X is dependent on the existence of a correspondence between states of v and states of X. Cases where this correspondence does not hold are regarded as *misrepresentations*. For example, we might say that a neuron represents the color RED if its firing rate corresponds to instances of a red stimulus. A high firing rate in response to an instance of red is therefore a *correct* representation of RED. And that same firing rate *without* an instance of red is, necessarily, a *misrepresentation* of RED. In other words, if this neuron exhibits a

[12] At least among philosophers. A recent study by Favela & Machery (2023) suggests neuroscientists do not necessarily regard intentionality in this manner.

high firing rate in response to *anything other than RED*, then that is a misrepresentation. In general, this shows that regarding a physical vehicle v as a representation of X not only defines the relation between states of v and states of X (as a *correct* representation) – it also necessarily defines the relation between these states of v and *anything other than X* (as an *incorrect* representation). That is the mark of an exclusive relation. Just as Alice being "legally married to" Bob doesn't just define the relation between Alice and Bob (as being legally married), but also the relation between Alice and *anyone other than Bob* (as *not* being legally married).

These characteristics of representation have long been discussed by philosophers and are often described as the "determinacy of content". Yet we will refrain from using this term. For one, we think "exclusivity" better captures the relevant feature of the representation relation. But also, "content determinacy" is regularly contrasted with the possibility of disjunctive contents, while "content exclusivity", as we understand it, is consistent with disjunctive contents. That v is a representation of *only* X, does not mean that X itself cannot be a disjunction. We think that the contents of representations in general, and neural representations in particular, can be disjunctive.[13] Yet, they are still exclusive. A representation of $(X \vee Y)$ is a representation of *only* $(X \vee Y)$.

Notably, a representation of $(X \vee Y)$ is necessarily *not* a representation of X, as well as necessarily not a representation of Y. This is evident by considering the truth conditions of such representations. If a certain neural state carrying the disjunctive content "RED or LEFT" responds to a stimulus which is *green* and moving left, then that would be a *correct* representation. As such, this neural state is necessarily *not* carrying the content RED. A parallel point can be made regarding conjunction. A representation of $(X \wedge Y)$ is necessarily *not* a representation of X, as well as necessarily *not* a representation of Y. Taken together, this further illustrates the exclusivity of *any* content: If a representation of $(X \vee Y)/(X \wedge Y)$ is necessarily not a representation of X, then the same is true the other way around – a representation of X is necessarily not a representation of $(X \vee Y)/(X \wedge Y)$, for *any non-coextensive property* Y. This is far from a trivial characteristic for a type of relation. For example, if v has the property of *being* RED, then *as such*, it likely has the property of *being* $(RED \wedge something\ else)$, and it certainly has the property of *being* $(RED \vee something\ else)$. Yet, if v is a *representation* of RED, then as such it is necessarily

[13] The go-to example in these discussions is usually the frog's fly catching mechanism (Lettvin et al., 1959) and whether, in catching flies, the frog's relevant inner state represents "flies" or "food" or "small, dark, moving object". In the context of this paper, we are happy to accept that the frog might represent any disjunction of these (or other) possibilities (see Millikan, 2023).

not a representation of $(RED \wedge something\ else)$, and necessarily *not* a representation of $(RED \vee something\ else)$. As a representation of X, v is a representation of *only* X.

3.2 Reconciling Content Multiplicity with Content Exclusivity

Having come to grips with the exclusivity of the representation relation, its tension with the possibility of content multiplicity becomes apparent. If v representing X means that v is representing *only* X, and v representing Y means that v is representing *only* Y, then how could v represent *both* X and Y? How can v represent two (or more) different things *while still maintaining the necessary exclusivity of the representation relation*?

To answer this question, we can begin by thinking of the simple "legal marriage" example. We ask: if v being "legally married to" X means that v is "legally married to" *only* X, and v being "legally married to" Y means that v is "legally married to" *only* Y, can v be "legally married to" both X *and* Y? Well, there is a way this is possible – if we consider "legal marriage" as defined only relative to a specific country, v *can* be "legally married to" both X and Y *while still maintaining the exclusivity of the "legal marriage" relation*. There can be two different countries A and B, such that:

- Relative to A, v is "legally married to" only X (and not Y).
- Relative to B, v is "legally married to" only Y (and not X).

This way, the "legal marriage" relation remains exclusive, and multiplicity is possible because of the localization of each relation to a different country. This illustrates a more general truth – *the only way exclusivity and multiplicity can coexist is by "localizing" each exclusive relation.*

Take the "rabbit-duck" sketch, which was mentioned in section 1 as one of the most famous examples of content multiplicity. There, the "localization factors" A and B, correspond to a given observer at a given moment, such that:

- Relative to A, the figure represents only a rabbit (and not a duck).
- Relative to B, the figure represents only a duck (and not a rabbit).

To account for the different contents, A and B must of course be different. In the current example, this means either two different observers at the same moment, a single observer at different moments, or different observers at different moments. Either way, the multiplicity of "localization factors", A and B, enables the multiplicity of content. And within each localization, the exclusivity of the representation relation is maintained. In

fact, a given observer's inability to perceive both the rabbit and the duck simultaneously is, we believe, another nice illustration of content exclusivity.

To clarify, the point we wish to make is about localization, not observers. For the famous rabbit-duck sketch, observers act as the localizing factors, but that is just one example. In the next subsection, we will discuss other possible localizing factors. Crucially, what we claim here is that *some* localization is necessary. If v is a representation of two distinct contents X and Y, then there must be *some* localizing factors A and B, such that:

- Relative to A, v is a representation of X (and only X).
- Relative to B, v is a representation of Y (and only Y).

Importantly, this means that the only way to account for content multiplicity is to accept that localization plays a *constitutive* role in determining the representational content. In particular, what enables this account of content multiplicity is the understanding that each exclusive content is only defined relative to exactly *one* localizing factor, *and not the other*.

It would perhaps be helpful to consider the "legal marriage" example once more. Suppose there are two countries A and B such that, relative to A, v is legally married to X, and relative to B, v is legally married to Y. What would happen if the two countries were to unite to a single country "A and B"? v would now be in a bind. Assuming the exclusivity of the "legal marriage" relation must be maintained, it follows that, relative to "A and B", v is *not* "legally married" to X, as well as *not* legally married to Y. The localization to A, *and A alone*, was necessary to define v's "legal marriage" to X, and without it the relation does not exist. Likewise, the localization to B was necessary to define v's "legal marriage" to Y.

The same logic must apply for any exclusive relation, including representation. If v is a representation of X relative to A, and a representation of Y relative to B, then A and B are the localizing factors that account for the existence of two different exclusive relations. Considering "A and B" *together* undermines the localization that enables this account. v's representation of X is only defined in virtue of the localization to A *alone* (and particularly, without B), while v's representation of Y is only defined in virtue of the localization to B *alone* (and particularly, without A). Without this localization, there would be no exclusive relation, and hence no representation.

3.3 The Problem for the Naturalistic Approach

Initially, our conclusions above might not seem all that problematic for the proponents of original contents. Perhaps they can accept the necessity of localization, as long as the

localization factors are themselves naturalistic. And indeed, it seems that existing naturalistic theories of content have obvious candidates for such localization factors. As mentioned in section 1, naturalistic accounts define representation by appealing to a variety of possible factors, besides the existence of a tracking relation. For example, many accounts appeal to some notion of teleological function (e.g., Dretske, 1988; Neander, 2017; Shea, 2018), and perhaps that can help provide the necessary localization factors to account for content multiplicity. If there are two different (natural, objective) teleological functions A and B, which are naturally differentiated by, say, two distinct evolutionary selection processes, then it would seem that each can define a distinct representational function for the same neural state v such that, relative to the teleological function A, v represents X, and relative to the teleological function B, v represents Y. So wherein lies the problem?

Well, the issue is with *the need for localization at all*, and not with which conditions act as localization factors. It is not that a particular localization factor contradicts the naturalistic accounts, but rather the in-principle idea of localization having a constitutive role in defining representational content. In other words, the problem is that *Mother Nature doesn't really do localizations*.

Suppose, for example, that A and B are indeed two natural, objective, teleological functions, and one wishes to claim that:

- Relative to function A, v is a representation of X (and only X).
- Relative to function B, v is a representation of Y (and only Y).

As was stressed in section 3.2, this means that v's representation of X is necessarily dependent upon the localization to function A *alone* (and particularly, without B), which is constitutive to v's representation of X (and similarly, the localization to function B *alone* is constitutive for the representation of Y). But crucially, there is no naturalistic justification for this localization to each function *alone*. Nature might "give us" the two teleological functions A and B, but it does not choose each one *on its own*. It does not give us *A without B*, or *B without A*, only *A and B as a pair*. And yet, it is precisely the localization to *only A* (and *only B*) which is necessary to define v as a representation of X (and of Y).

One might attempt to push back on this claim by noting that if there are two distinct teleological functions, then there must be two distinct selection processes that defined them. Each selection process defines exactly one function, and not the other. That way, one might claim, nature does give us each function *on its own*. But such a claim does not solve the problem for the naturalistic approach, it merely moves it along. Having

distinct selection processes is not enough. We already know that the two functions are distinct, the problem is that there are *two* of them. Similarly, even if each function has its own distinct selection process, there are still *two* of those. The focus on exactly *one* process, and not the other, to define exactly *one* function, and not the other, is *an implicit choice*. There is nothing in nature that justifies this choice, nothing which justifies considering each one process while excluding the other.

Similar considerations can be applied to other possible naturalistic localization factors. For example, the same neural vehicle v can have multiple downstream effects, and multiple causal roles. Naturalistic theories can appeal to these distinct effects to act as localization factors and define distinct contents.[14] And there can be other possibilities as well (see section 1). It is likely that most naturalistic theories will be able to provide distinct natural conditions to act as localization factors, but this would not suffice to account for content multiplicity, for the same reason discussed in the previous paragraph. Namely, even if we have two distinct natural, objective conditions A and B to localize representations X and Y, there is no naturalistic justification to *consider one without the other*. If A and B are both naturally occurring conditions, nature itself *does not* choose each one of them *on its own*. And yet, as discussed in 3.2, that is precisely what we need to account for each exclusive content. Regardless of what the precise localization factors are, the same issue will continue to plague any naturalistic attempt.

3.4 Subjectivity to the Rescue

To recap, in section 3.2 we showed that to account for content multiplicity each exclusive content must be localized, and in section 3.3, we saw that while there might be a number of relevant natural localization factors, there is no naturalistic justification for localization in itself. How then, can a theory of neural representations provide the localization necessary for content multiplicity?

As far as we see, the only way to bridge this gap is by appealing to pragmatic accounts of neural representation. As mentioned in section 1, on pragmatic accounts the researchers' focus on particular (natural) facts plays a constitutive role in defining the contents of neural representations (Egan, 2014, 2018; Cao, 2022; Hacohen, 2022). And it is precisely this appeal to the current interests and choices of researchers which enables the necessary localizations in cases of content multiplicity since it provides a way for

[14] In goal-directed or consumer-based accounts of teleosemantics (e.g., Millikan, 1984, 1989), this can coincide with the option of appealing to teleosemantic functions as localization factors.

choosing each localization factor *on its own*. That is, subjectivity is not needed to define the localization factors themselves, but rather to provide a means for *favoring between them*.

As discussed in section 3.3, nature is not capable of providing a way to consider each localization factor *on its own*. But *we*, external cognitive agents, can. *We* can define v as being legally married to *only X*, as well as being legally married to *only Y*, by considering each country independently. *We* can view the rabbit-duck sketch in Figure 1 as a rabbit, as well as a duck, by accepting different interpretations of its features. And *we* can consider the prefrontal population from Mante et al. (2013) as a representation of COLOR, as well as a representation of MOTION. We do this by focusing on distinct localization factors, such as orthogonal coding dimensions or different aspects of the task. The localization factors themselves can be defined naturalistically, but it is the researchers' *subjective* focus on each one factor *on its own* that enables the localization which is necessary to define each exclusive representational content.

This leaves a relatively modest role for the subjective considerations of researchers in defining the contents of neural representations, but it is a *constitutive role nonetheless*. It follows, therefore, that these representations are dependent on the intentions of external cognitive agents and *do not carry original contents*.

4 Conclusion

In this paper we have argued for two claims:

1. Content multiplicity is inherent to the notion of neural representation that is used in contemporary neuroscientific practice (section 2).
2. Naturalistic theories positing intrinsic or original contents, cannot account for content multiplicity, while pragmatic theories can (section 3).

It is worth noting that each of these two claims is independent of the other, and both merit consideration in their own right. Understanding how content multiplicity figures into contemporary neuroscientific practice (section 2) offers valuable insight into the nature of neural representations, regardless of the argument in section 3. Similarly, our argument that content multiplicity cannot, in principle, be accounted for by representations with original contents (section 3), places a significant limitation on naturalistic theories of content, regardless of the argument in section 2. When taken together, the two claims amount to show that the notion of representation that is used by contemporary

neuroscience cannot be accounted for by naturalistic theories, and that neural representations do not carry original content.

One might argue this conclusion only applies to neural representations that exhibit content multiplicity. However, we believe that qualifying our conclusion in this manner would be ill-advised. For one, as discussed in section 2, neuroscientific practice is in the process of a general paradigm shift towards population coding, and thus, examples of content multiplicity will quickly become the new norm.[15] Hence, drawing the distinction between neural representations that allow for content multiplicity, and neural representations in general, is misguided. Allowing for content multiplicity should be regarded as an essential characteristic of any theory of neural representations.

Moreover, even if examples of content multiplicity were scarcer, it would still likely be unjustified to posit the existence of two inherently different notions of representation. If one accepts that the contents of neural representations in cases of content multiplicity are necessarily *not* original, what justifies concluding that in other cases they are? We should strive to understand the notion of representation that is used in neuroscience in a unified manner. If a pragmatic theory of representation is the only way to successfully account for the explanatory role of neural representations in cases of content multiplicity, it is likely to be the right way to account for that same explanatory role in other cases as well.

[15] And there is a case to be made that single neuron coding also exhibits content multiplicity (e.g., Gawne, 2000).

References

Aoi, M. C., Mante, V., & Pillow, J. W. (2020). Prefrontal cortex exhibits multidimensional dynamic encoding during decision-making. Nature neuroscience, 23(11), 1410-1420.

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. Nature reviews neuroscience, 7(5), 358-366.

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. Trends in Cognitive Sciences, 26(11), 942-958.

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

Brentano, F. (1874 [1995]). Psychology from an empirical standpoint. London: Routledge.

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

Cao, R. (2022). Putting Representations to Use. Synthese, 200(2), 1-24.

Cunningham, J. P., & Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. Nature neuroscience, 17(11), 1500-1509.

Chung, S., & Abbott, L. F. (2021). Neural population geometry: An approach for understanding biological and artificial neural networks. Current opinion in neurobiology, 70, 137-144.

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

Dretske, F. (1995). Naturalizing the Mind. Cambridge, MA.: MIT Press.

Ebitz, R. B., & Hayden, B. Y. (2021). The population doctrine in cognitive neuroscience. Neuron, 109(19), 3055-3068.

Egan, F. (2014). How to Think about Mental Content, Philosophical Studies 170(1), 115-135.

Egan, F. (2018). The Nature and Function of Content in Computational Models, in M. Sprevak and M. Colombo (eds.), The Routledge Handbook of the Computational Mind. Routledge, 247-258.

Elber-Dorozko, L., & Loewenstein, Y. (2023). Do retinal neurons also represent somatosensory inputs? On why neuronal responses are not sufficient to determine what neurons do. *Cognitive Science*, *47*(4), e13265.

Favela, L. H., & Machery, E. (2023). Investigating the concept of representation in the neural and psychological sciences. *Frontiers in Psychology*, *14*, 1165622.

Fodor, J. A. (1987). Psychosemantics: The problem of meaning in the philosophy of mind. Cambridge MA: MIT Press.

Fodor, J. A. (1990). A Theory of Content and Other Essays, Cambridge, MA: MIT Press.

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: high dimensionality for higher cognition. Current opinion in neurobiology, 37, 66-74.

Gallistel, C. R. (1990). Representations in animal cognition: An introduction. *Cognition*, *37*(1-2), 1-22.

Gawne, T. J. (2000). The Simultaneous Coding of Orientation and Contrast in the Responses of V1 Complex Cells. *Experimental Brain Research* 133 (3):293–302.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. Annu. Rev. Neurosci., 30, 535-574.

Hacohen, O. (2022). What Are Neural Representations? A Cummins Functions Approach. Philosophy of Science, 89(4), 701-720.

Hanks, T. D., & Summerfield, C. (2017). Perceptual decision making in rodents, monkeys, and humans. Neuron, 93(1), 15-31.

Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. & Pitts, W. H. (1959). What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers*, **47**(11), 1940–1951.

Mante, V., Sussillo, D., Shenoy, K. V., & Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78–84.

Millikan, R. (1984). Language, Thought and other Biological Categories. Cambridge, MA: MIT Press.

Millikan, R. (1989). Biosemantics. Journal of Philosophy, 86: 281–97.

Millikan, R. (2023). Teleosemantics and the frogs. *Mind & Language*, 1–9.

Neander, K. (2017). A mark of the mental: In defense of informational teleosemantics. MIT Press.

Piccinini, G. (2022). Situated neural representations: Solving the problems of content. Frontiers in Neurorobotics, 16.

Saxena, S., & Cunningham, J. P. (2019). Towards the neural population doctrine. Current opinion in neurobiology, 55, 103-111.

Shadlen, M. N., & Newsome, W. T. (1996). Motion perception: seeing and deciding. Proceedings of the national academy of sciences, 93(2), 628-633.

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

Vyas, S., Golub, M. D., Sussillo, D., & Shenoy, K. V. (2020). Computation through neural population dynamics. Annual review of neuroscience, 43, 249-275.

Wittgenstein, L. (1953). Philosophical Investigation. Oxford: Blackwell

Yuste, R. (2015). From the neuron doctrine to neural networks. Nature reviews neuroscience, 16(8), 487-497.