Final draft forthcoming in *Synthese.* Please reference the published version when available.

Introduction to the Topical Collection 'Locating Representations in the Brain: Interdisciplinary Perspectives'

Sarah K Robins [1] & Felipe De Brigard [2,3,4]

1 Department of Philosophy, Purdue University

2 Department of Philosophy, Duke University

3 Department of Psychology and Neuroscience, Duke University

4 Center for Cognitive Neuroscience, Duke University

Final draft forthcoming in *Synthese.* Please reference the published version when available.

Introduction to the Topical Collection 'Locating Representations in the Brain: Interdisciplinary Perspectives'

We know that Twitter is not what it used to be, but if you were around the Twittersphere in 2019, you may remember a series of long discussions, by very well-known neuroscientists, on the nature of neural representation. Reading from the bleachers, many philosophers like us couldn't help but notice that some of the themes discussed in these threads were very familiar. Indeed, they were uncannily similar to the way philosophers of mind argued in the 1970s and 1980s about the prospects of naturalizing intentionality. While the recent debates were couched in terms of multivariate analyses, pattern similarity, and repetition suppression, they were ultimately about how to understand misrepresentation, representational content, and even reference to abstract and non-existent entities, albeit in the context of contemporary cognitive neuroscience. The time was then ripe to try to bring together both philosophers and neuroscientists interested in the nature of representation, so they could talk to and learn from each other.

A grant from the Summer Seminars in Neuroscience and Philosophy, which has been taking place at Duke University since 2015, as well as a conference supporting grant from the National Science Foundation, provided the funds for running an interdisciplinary conference at the Neuroscience Institute at Stanford University in the Fall of 2019. For three days, several philosophers and neuroscientists debated the difficult issue of how to understand the notion of *neural representation* in contemporary brain science, a construct that for many is foundational to cognitive neuroscience. Many of the speakers at this conference turned their papers into contributions for this topical collection, which seeks to bring philosophers and neuroscientists together in service of the larger goal of arguing constructively about what it means to fit representations in the brain. To help to navigate the contents of this topical collection, we start off

with a brief tour of some of the main themes in the philosophy of mind on the difficulties of naturalizing mental representations as well as a quick introduction on the history of neuroscientific research on the nature of representations in the brain.

1. Philosophy of Naturalizing Mental Representation

Appealing to mental representations to explain behavior and mental phenomena is a relatively recent development with a very old history. It is recent insofar as the term "mental representation" is employed as a technical term in contemporary philosophy of mind and cognitive science. At the same time, the notion behind the term has been present for a long time, in the work of philosophers that sought to explain thought in terms of mental particulars bearing names such as "images", "impressions" or "ideas". What all these terms had in common, though, is their use to refer to mental particulars standing in place of the things they were about. Instead of trafficking with actual trees, when the mind thinks of a tree it does so via an impression or an idea *of* a tree, whereby that "of" expresses a semantic relation. This precise semantic relation was famously characterized by Brentano (1874) as the "intentionality" of mental states: unlike physical phenomena, mental phenomena are about things. That which a mental state is about is known as the *content* of the mental state, and it is to be distinguished from its *object*, which need not exist. One can think about the Tooth Fairy, and such a thought would have content-–i.e., it would be about the Tooth Fairy– even though the object of the thought does not exist. Importantly, Brentano took intentionality to be "the mark of the mental" in a strong metaphysical sense: that is, he argued that intentionality was a property of mental states that made them essentially distinct from physical states. Psychology, understood as the science of the mind, could thus proceed independently of the physical sciences, to which intentionality could never be reduced.

The irreducibility of intentionality was further bolstered, over half a century later, by Chisholm (1957), who argued that sentences expressing intentional statements have linguistic properties (e.g., no existential generalization from seemingly referential substantival expressions; no replacement of co-referential expressions *salva veritate*) that makes them irreducible to non-intentional vocabulary. A few years later, however, having accepted the irreducibility of intentional statements into non-intentional ones, Quine (1960) presented philosophers of mind with a difficult dilemma: either to accept the irreducibility of intentionality and embrace the ontological consequence that the mental is *sui generis* and that the science of the mind, if there was to be one, could not be continuous with the natural sciences. Or to reject the irreducibility of intentionality and bear the burden of explaining how the mental can be understood in non-mental terms, so that the science of the mind, if there was to be one, could be continuous with the natural sciences. Those who pursued the second horn of the dilemma were engaged in the project of *naturalizing intentionality*, of trying to find a way to bridge the gap between our physiological and our psychological theories (Dennett, 1969).

With the advent of philosophical functionalism and the birth of cognitive science, a new *computational theory* of mind offered to help to bridge that gap. Thinking of mental processes in terms of computations carried by physical representations promised to offer an answer as to how intentional contents can, contra Brentano, be continuous with the natural world. But the project of understanding the nature of the physical representations over which mental computations operate has proven challenging. Theorists disagree, for instance, as to whether the vehicles of such representations should be local or distributed across the brain, or whether we should think of those representations at a personal or at a subpersonal level. But perhaps one of the hardest questions to answer for a naturalistic theory of mental representation, is the question about *content*

determination: roughly, how is it that a particular physical representation comes to mean what it does. Of course, we know how a number of physical representations acquire their meanings. The physical symbol ">" in a written mathematical formula is interpreted as "greater than" because at some point it was introduced to play that role in mathematics and ever since people have used it that way. These instances of "derived intentionality", however, depend on people's mental states, and as a result won't be able to help us explain how it is that physical brains like ours come to have representations with intentional contents. What we need is a theory of "original" or "primitive intentionality".

One attractive possibility is to think of content-fixing in terms of *causation*: if a token representation, *R*, is caused by and only by instances of property *P*, then the content of *R*, *Rp*, means *P*. Unfortunately, causation is not a good content-fixing relation, for at least two reasons. The first one pertains to the problem of *misrepresentation*. As it happens, sometimes *Rp* is triggered by things other than *P* –say, you may think you saw a snake when in reality it was just a twig. A theory of content should be able to explain why it is that sometimes mental states can misrepresent. A second reason pertains to *co-instantiation*: it often happens that a property *P* is co-instantiated with a property *Q*, so that *R* is triggered both when *P* is present as well as when *Q* is present. The problem is that *Rp* is not ambiguous between *P* and *Q*: I can think about an entity as cordata (i.e., having a heart) or I can think about an entity as renata (i.e., having a kidney), and these thoughts would have different content even though both renata and cordata are co-instantiated properties (Quine, 1960). Causation seems to be too coarse a relationship for the fine-graininess of mental representation.

A more popular possibility –indeed, more popular among neuroscientists-– is to cash out the content-fixing relationship in terms of *correlations*: if *R* correlates with *P* then *Rp* means *P*.

Unfortunately, this option won't work either, not only because it can't solve the problem of misrepresentation and the problem of co-instantiation, but also because it has a problem of its own: lots of natural states co-vary with other natural states without one having the other as its content. An attractive variation on this theme, inspired by Shannon (1948) and then Dretske (1981), was to think of the content-fixing relation in terms of *informational* correlation: if the tokening of *R* increases the probability of there being *P*, then *Rp* means *P*. Alas, once again, problems abound. Not only because it isn't clear whether this informational correlation view can account for the problem of misrepresentation *tout court*, but also because it faces other difficult problems, such as the problem of *implication* and the problem of *disjunction*. Suppose that you token a belief the content of which is that *p*, and suppose further that *p* implies *q*. As a result, if believing that *p* increases the probability of *p* being the case, then it should also increase the probability of *q* being the case. But of course you may not believe that *q*. So increased probability alone won't do. The problem of disjunction is related. Often, the targets of our representations are distal. The activation of a particular neuronal population in the presence of a black spot in the visual field of a frog may increase the probability of there being a nearby fly. But it is *more* probable that there is a fly *or* an eagle. However, the content of the representation is not disjustive, but determined. So it looks like mere increased probability, again, won't do.

In the last three decades there have been several attempts to try to solve these and related concerns posed by the challenge of naturalizing intentionality in mental representations. Among them, two of the most influential are the Asymmetric Dependency Theory (ADT, Fodor, 1990) and Teleological theories (Dretske, 1988; Millikan, 1984; 1989; Neander, 2017). A good starting point to understand ADT, is to ask–along with Fodor (1986)—whether paramecia have mental representations. There are plenty of organisms whose behaviors are nomically, statistically and/or

informationally associated with certain properties in the environment. The rings in tree trunks, for instance, are associated with seasonal changes and, thus, can indicate the passage of time. But it would take some doing to say that they *represent* time. Likewise, paramecia show avoidance reactions whenever their outer cilia bend beyond a particular threshold against a strong surface. Interestingly–and this is something Fodor likely didn't know–the internal mechanism that produces the aversive reaction in paramecia is the exact same voltage change that occurs in an action potential, which is why paramecia are sometimes called "swimming neurons" (Brette, 2021). Yet, despite this *ceteris paribus* nomic association, we don't say that paramecia can represent obstacles. The reason, according to Fodor, is that in the case of paramecia, the *ceteris paribus* nomic relation that governs the internal state's reaction to the external stimulus is obligatory, so paramecia can't be selective as to whether to respond to the relevant stimuli. Moreover, paramecia cannot respond to external stimuli for which there *isn't* a nomic relation with an internal state. By contrast, organisms that represent, like us, can respond to external stimuli for which *no* nomic connection exists, and also fail to respond to external stimuli for which there *is* a nomic relation with an internal state. Nevertheless, in both organisms, the relevant nomic relation still features in the explanation of their behavior.

To account for this observation, Fodor suggests understanding the content-fixing relation in terms of ADT. According to his view, *Rp* means that *p* if and only if there is an asymmetric causally dependent relationship between the presence of *p* and the tokening of *Rp*. The thought, in essence, is that while (*ceteris paribus*) *p* causes *Rp*, there are instances in which *Rp* is caused by non-*p* things—say, *q*. However, when this occurs, it only happens because there is already a nomic association between the presence of *p* and the tokening of *Rp*. Had there not been a nomic association between *p* and *Rp*, then *q* wouldn't have token *Rp*. Here's an example. Suppose I've

learned to identify copperheads so that whenever I see a specimen of the venomous species *Agkistrodon contortrix* (i.e., a copperhead), I entertain the concept COPPERHEAD. One day, though, walking in the woods, I token the concept COPPERHEAD in the presence of an Eastern Water Snake, i.e., a specimen of the harmless species *Nerodia Sipedon*. This is a clear instance of misrepresentation, and a very common one, given how similar both specimens are. The ADT handles this case by pointing out that the reason why I misidentified the eastern water snake as a copperhead is because there is a nomic causal relation between true instances of COPPERHEAD and the presence of copperheads. Had there not been this causal relationship to begin with, my case of misidentification wouldn't have been a case of misrepresentation.

Unfortunately, the ADT is problematic as well. Some have pointed out, for instance, that ADT cannot handle cases of uninstantiated properties (e.g., UNICORN) as there is no obvious reason for *Rp* and *p* to be the privileged nomic relationship rather than the relation between *Rp* and *q*. Without a principled way of establishing nomic privilege, the asymmetry breaks down (Baker, 1991).[1] Another criticism concerns the fact that ADT cannot solve the problem of co-instantiation, for if the property in virtue of which *p* causes *Rp* is necessarily co-instantiated with another property, say *q*, then the asymmetry in the nomic relationship once again breaks down (Gates, 1996). But perhaps the most serious worry with ADT is that this putative relation of asymmetric causal dependence sounds thoroughly normative and, as a result, it seems very hard to spell out in purely descriptive and naturalistic terms (Lower, 2017).

The second alternative, Teleological theories, has the advantage of tackling the concern about naturalizability head on, for according to this view, the content-fixation relation is

[1] Notice, incidentally, that this is exactly why for the ADT a paramecium wrongly avoiding a large piece of food as if it was a threat does not constitute a case of misrepresentation: the cilia would have reacted avoidantly regardless of whether what it encountered was a larger than usual piece of food or a typical obstacle. The worry is that the same applies to some bona fide cases of representation.

established by our own biology. Specifically, in its more general form, teleological theories of mental content hold that, for an organism *O*, *Rp* means *p* if (and, for some versions, only if) *R* serves the function of carrying information about *p* for *O*. In its classical formulation, the notion of function was etiological, so that the function of a particular *R* was to be cashed out in terms of why, historically, tokens of that type were selected for by evolution (Millikan, 1989; Neander, 1991). Additionally, teleological theories distinguish between the way in which a representation is *produced* in the organism from the way in which it is *consumed* by the organism, as the latter determines the content-fixing relation. Consider paramecia again. A single paramecium manages to avoid threatening obstacles when its cilia react to certain conditions, such as changes in local currents, which are typically associated with foreign bodies in its immediate vicinity. Voltage-gated L-type calcium channels (just like those found in neuronal synaptic terminals) enable the transduction of an ionic current, which in turn triggers an action potential. For the duration of the voltage change the paramecium swims backwards, effectively avoiding the obstacle and returning back to its baseline helicoidal navigation. The disturbances in the paramecium's surrounding milieu are typically caused by threatening obstacles, so the signal produced by the cilia, which in turn causes the action potential that enables the organism to avoid the threat, is consumed (or used) by the paramecium as a sign for the presence of a threatening obstacle. The organism qua consumer of the signal determines the content of the representation as that of a threatening obstacle. Notice, then, that on the teleological view, the assessment of paramacia's avoiding behavior would be radically different from Fodor's, as now the microorganism *does* appear to make use of representations. Moreover, the theory handles the problem of misrepresentation in paramecia easily: when the mechanism is triggered by non-threatening obstacles, such as food, then the paramecium is consuming the signal the way it *normally* would, whereby 'normal' is to be

interpreted in accordance with its natural function. The misrepresentation occurs because the content-fixing relation is given by how the organism consumes the representation, rather than how it is produced.

The consumers of a representation need not be organisms as a whole. Subsystems within an organism can be consumers too. This clarification allows us to see how one could apply the teleological approach to representations in the brain. Here's an illustration. In Hubel and Wiesel (1959) famous study, particular neurons in the striate (visual) cortex of the cat were shown to selectively respond to specific orientations of a slit of light. Indeed, further studies showed that neighboring neurons in BA 17 responded to different orientations, each of which could be matched to corresponding orientations in the activated retinal ganglion cell, which in turn could be mapped onto stimuli in the cat's visual field. These signals, in turn, are consumed by the cat's motor system to fine-tune movements accordingly. Here, the motor system consumes the signal produced by the neurons in the striate cortex, thus enabling us to say that they represent orientations. Moreover, this account allows us to explain why, under certain circumstances, such as unusual lighting or peculiar color-contrasts, the motor system can fine-tune a motor response as if a particular orientation was present to the visual system when in reality it was not. If you've ever tripped on a dimly lit escalator because you didn't see where the step was emerging, you've experienced this case of misrepresentation.

Unfortunately, though, the literature is also plagued with concerns about the explanatory limits of teleological theories of representational content. For one, it has been argued that they can't handle some cases of content underdetermination due to co-instantiated properties (Fodor, 1990). Others are concerned with the theory's reliance on the evolutionary history of representational systems. Suppose a replica of a representational organism is artificially created *de*

novo, lacking thus any evolutionary connection to the history of selection that produced the representational mechanisms in the original organism. Intuitively, this artificial organism would be representational as well, yet a classical teleological theory would have to say that it isn't. Others complain that the theory is too liberal, granting representational capabilities to entities that likely don't have them, such as plants and, as we saw, paramecia. And finally, and perhaps more critically, it is very hard to see how this view scales up from basic, low-level representations-– such as colors, edges, or even basic reinforcements or rewards–to complex concepts and elaborate thoughts. Even the most developed versions of teleological accounts in the offing admit of their own limited explanatory power, as they are confined to either non-conceptual (Neander, 2017) or sub-personal representations (Shea, 2018).

There is no need to expand this brief overview of the main philosophical approaches to naturalizing intentionality in mental representation because, as Lower (2017) reminds us, the take home message is clear: the more naturalistic the theory is (e.g, causal, correlational, teleological), the less it accounts for the complex nature of intentional content, whereas the more the theory seems to account for such complex intentional contents (e.g., ADT), the less naturalistic it seems to be. And our view is that this very same message should not be ignored by neuroscientists interested in understanding how brains like ours manage to represent. As we will see in the next section, neuroscientists employ the term "representation" ubiquitously. However, what they mean by it is often unclear, and when it is clear, it often means something like causally or correlationally related to a particular stimuli, typically an existing one. As such, even when neuroscientists are straightforward about their use of the notion of representation, which is unusual, their meaning is confined to one of those uses that philosophers for decades have found, if not problematic, at least severely limited. Moreover, as some of the articles in this special issue make it clear, some uses of

the term "representation" in neuroscience are surprisingly unnaturalized, as scientists often sneak in all sorts of normative considerations when constraining their stimuli set, their model's parameters or their analytic strategies. Since we are all in the business of understanding how brains can represent, we believe the philosophers' concerns should not be taken lightly.

2. Neuroscience: History of Locating Representations in the Brain

The idea that the brain represents things also has a venerable history. For almost 15 centuries, the ventricular-pneumatic doctrine placed the functions of the mind in the ventricles of the brain. According to this view, the vehicle of our vital operations were pneumatic spirits, which flowed through the body via our circulatory system, and were ultimately stored in the ventricles, i.e. the spaces in between the two hemispheres of the cerebrum. The work of the anatomists of the Renaissance managed to debunk the idea that the ventricles were the seat of the mind and that the veins and arteries were the conduits of our sensations and movements. They showed instead that the brain itself was critical for mental function and that sensations and movements depended, not on our circulatory, but on our nervous system. Yet, the pneumatic portion of the theory remained for another two centuries, forcing natural philosophers and physicians to think about how animal spirits–which were thought closer to fluid or air than to solid matter–could send messages to and from the brain.

The solution, as it happens, was an earlier notion of a brain representation. In his *Treatise on Man*, for instance, Descartes (1667) talked about spirits passing through pores in the brain, the precise pattern of which represented different objects. In fact, the idea that different arrangement of pores is how brains represented objects was apparently so widespread during the 17th century, that it was included as a truism in the *Lexicon Medicum Graeco-Latinum*, a medical dictionary of

sorts published in 1684. Interestingly, by the time the pneumatic theory was abandoned, in part thanks to the work of early electrophysiologists such as Galvani (1737-1793), Legallois (1770-1840) and Flourens (1794-1867) who show that it was electricity rather than spirits flowing through the nerves what was responsible for sensations and motor actions, the idea that the brain represented both objects and movements still perdured. In J.A. Paris' influential *Pharmacologia* (1831), a medical textbook that saw many editions during the 19th century, the idea that the brain represents movements and sensations was so entrenched, that it was even used in explanations as to why, for instance, amputated patients continued to feel pain in limbs they had no more–what is known today as "phantom pain" (John, 2021).

By the mid-19th century, the idea that the brain was able to represent was no longer difficult to accept. The challenge, instead, was to understand how and where those representations were located in the brain. The quest became one of understanding how different research questions and tools could be used to explore different ways of thinking about what the brain is representing and how. Providing a comprehensive account of these challenges and the ways they have arisen in the centuries since would be a tremendous task, which we do not attempt here. Instead, we offer a brief survey of illustrative cases and places where debates about the nature of representations have been salient.

One approach to the brain's representational capacities begins from system-wide considerations: are the brain's functions distributed throughout or are functions found in dedicated regions or substructures? Proponents of the former are described as supporting *holism* or *equipotentiality*, while proponents of the latter are described as supporting *specialization* or *localization*. There have been numerous cycles of debate between these two positions over the last two centuries of neuroscience (see Mundale, 2002 and McCaffrey, 2022 for compelling reviews).

A key early iteration pitted the localist phrenology of Franz Joseph Gall (1757-1828) against the equipotentiality of Pierre Flourens (1794-1867). Although phrenology's reputation has been more enduring, in fact Flourens' was more influential amongst early 19th century brain scientists (Finger, 2002).

In the latter half of the 19th century, the view of the brain as organized by functional localization grew more dominant (Ward, 2023). Jean-Baptiste Bouillaud (1796-1881; Stookey, 1963) advocated for the localization of speech, a view ultimately championed by Paul Broca (1863).[2] The idea that the brain represents movements in "motor centers" as well as sensations of external objects in "sensory centers" was well received by the time W.J. Dodds wrote his "On the localization of the function of the brain", in 1878, and permeated even manuals of medicine popular by the end of the 19th century (e.g., *A dictionary of psychological medicine*, 1892, D. Hack Tuke, Ed.). As Ward illustrates in her review of the debate over the nature of motor representations during this period, there was broad consensus that the brain was in the business of representing and that it did so in particular ways. Ward notes, "all parties to the debate shared several operative assumptions about representation. Most notably, participants on both sides acknowledged that representations come in degrees and may overlap with one another" (2023: p. 16).

Karl Lashley (1890-1958) was a key figure in the return to equipotentiality in the early 20th century, as his exhaustive[3] search for the engram (a localized memory) failed. His work quieted research interest in identifying the neural correlates of particular memories (Robins 2023), indicating that memory performance was correlated not with any specific portion of cortex, but

[2] See Finger (2000) ch. 10 for an engaging discussion of Broca's role in these debates over cerebral localization.
[3] Lashley's search for the engram was restricted almost entirely to the cortex. As work in the subsequent decades would reveal, subcortical structures (especially the hippocampus) play a key role in memory storage.

rather with the amount of cortex involved. The discovery of selective memory deficits corresponding to the loss or degeneration of specific brain regions tipped the scales back toward localization. For example, Brenda Milner's (Scoville & Milner, 1957) work with HM and other patients identified the medial temporal lobe as critically involved in declarative memory. Here again the debate is between views of how the brain represents information - as distributed throughout cortex or localized in particular areas - not whether it is representational. Further support for localization came from Penfield's neurosurgical work, which demonstrated that memory-like states could be elicited from electrical stimulation of the lateral superior temporal lobe (Penfield & Perot 1963).

The emergence of neuroimaging techniques, PET and fMRI, in the 1980s and 1990s initiated a new period of interest in functional localization, spurring a new round of debate. Early work generated a great deal of excitement regarding the ability to identify neurocognitive systems, where specific cognitive functions could be mapped onto discrete neural structures (as reviewed by Viola & Zanin 2017). Others have decried this work as the "new phrenology" (Uttal 2001) and many since have systematically investigated evidence of neural reuse (Anderson 2010) and the multifunctionality of brain regions (McCaffrey 2015). Yet again, the debate concerns *what* not *whether* the brain is representing: for example, is the fusiform gyrus dedicated to representing faces (Kanwisher et al. 1997) or does it process representations of any specialized domain of visual information (Gauthier et al 2000)?

While answers vary, a shared methodological assumption remains: the answer will be determined by establishing clear relations between neural activity and content-specific stimuli and/or content-specific effects. The debate thus continues in other domains as new tools are brought in to establish or challenge these connections. For example, using TMS to stimulate the

brain regions central to debates about face representation shows selective impairment to some aspects of face perception and not others—i.e., the ability to recognize expressions, but not the ability to identify faces (Pitcher et al. 2008). Similarly, MVPA techniques are now often used to probe subtler statistical dependencies between patterns of neural activation and patterns in experimental stimuli that are detectable through established categories of brain regions and stimuli (Haxby et al. 2014).

A second approach to identifying representations in the brain takes place at a much finer level of grain: in the activity of single neurons. This approach to neural representation got a somewhat slower start than the work discussed above in the holism vs. localism debates. Indeed, the term "neuron" was not coined until 1891, by Wilhelm von Waldeyer (Finger 2002). The emergence of the neuron doctrine is most closely associated with the work of Santiago Ramón y Cajal (1852-1934), in deference to his thorough and influential advocacy for the neuron as the brain's primary functional unit.

The subsequent decades involved steady inquiry into the mechanisms by which neurons acted and interacted, which were increasingly characterized in terms of representing and transmitting information. Using the vacuum tube to amplify signals from implanted electrodes, Edgar Adrian (1889-1977) developed a key method for recording electrical signals in the nervous system. Correspondingly, he offered a view of nerve signals as carrying information. This information was not internal to the nerve signal, but rather was transmitted by the rate of signaling, like a code. Adrian's influential theorizing about how information was carried in the nervous system was deeply indebted to the communicative tools of his era (Garson, 2015). Later advances came through the exploration of the ways in which the signaling of neurons in specific parts of the brain is dedicated to particular features of the environment. Hubel and Wiesel (1959; 1962) work

on the details of information-processing in the visual system, discussed above, is a key example. They demonstrated that neurons in the visual cortex were not only selectively responsive to visual information, but also that individual neurons within these cortical areas were tuned to particular kinds of visual information - e.g., orientations of light-dark gradients.

Cognitive map theory (O'Keefe & Nadel, 1978) offers an account of hippocampal function based on what the neurons in this structure are representing: space and its relations. This view of the hippocampus is derived from the differential response profiles of distinct cell types in this structure: place cells (O'Keefe & Dostrovsky, 1971), grid cells (Hafting et al., 2005), border/edge cells (Sostad et al., 2008), etc. Collectively, they provide a map representing the animal's relative and absolute position.

As in the debates between localism and holism about mental representation at the macroscale, debates over neural representation at the unit level are also generally debates about what is being represented, not whether there are such representations. From the opposite end, we see how questions of content determination arise in neuroscience. Even once the pairing between specific stimuli and selective neural activity is well-established, yielding consensus as to what is being represented, questions remain about how to characterize the representation. For vision, this is most often illustrated through Lettvin, Maturana, McCulloch, and Pitts paper "What the Frog's Eye Tells the Frog's Brain" (1959) where the question of how to identify and label the stimulus remains. For place cells, debate continues as to their pairing with proximal or distal cues, ego vs allocentric representation of place, and even whether they are truly selective for spatial information exclusively (e.g., Aronov et al, 2017). There is, however, a general commitment to the idea that a neuron represents a particular content because of its correlation or covariance with a particular stimulus or feature of the environment.

Neural network models continue to view individual neurons as the key drivers of the brain's representational capacities, expanding on the general idea of neural representations coming about from population-level dynamics. On such a view, information is represented through patterns of activity that are distributed across neural populations. Such a system allows for the same large population to represent a range of distinct informational states within the same network - each as a distinct pattern (McClelland et al, 2014). Such approaches stand in stark contrast to the idea of a 'grandmother neuron' - a term invoked by Lettvin in a 1969 lecture, with the intention of ridiculing the idea of discrete, selective representation of information by a particular neuron or neurons. The idea nonetheless gained traction and plausibility for many - and even if literally false, proved valuable to the overall inquiry into selective neural representation (Barwich, 2019).

Network models of neural representation have grown increasingly influential, as machine learning techniques are used to identify abstract and high-dimensional representational patterns from measures of neural activity. Different models of hierarchical learning are used to approximate neural function and infer what the intermediary forms of neural representation might be. Such work forms the core an emerging and increasingly popular approach to neural representation, recently coined as the "neuroconnectionist research programme" (Doerig et al, 2023). The approach is intended to be situated at the intermediary point between high-level systems approaches to cognition and low-level mechanistic models of biological function. By including just enough neural plausibility, coupled with the power of abstract models, the hope is to have identified the ideal level at which to characterize what the brain is representing and computing.

In short, the history of locating representations in the brain is less a story of when and why neuroscientists decided to start using terms like representation to characterize the brain and more a story about how different research questions and tools have been used to explore different ways

of thinking about what the brain is representing and how. Iterations of it show up in system-level debates about functional localization, disputes over how to interpret the effects of discrete manipulations of neural circuits at the cellular level, and concerns over the use of decoding and other statistical techniques to analyze neural activity.

3. This topical collection: Interdisciplinary Perspectives

This topical collection starts with a provocative paper by Russell Poldrack, from Stanford University. The paper begins by acknowledging the indisputable fact that neuroscientists are all too comfortable using the term "representation" roughly to refer to "a systematic relationship between features of the natural world and the activity of neurons in the brain" (p. 1308). However, traditional and novel strategies to identify such systematic relationships–from single cell recording to encoding models to representational similarity analyses–suffer from a number of shortcomings, and are unlikely to be the last word on how to conclusively establish structural isomorphism between neural representation and their contents. Nevertheless, argues Poldrack, recent work in artificial networks –in particular, hierarchical convolutional neural networks (or HCNN for short)– offer a promising strategy to better understand the systematic connections between neural activity and the features of the world they represent. In particular, he argues that the hierarchical architecture of HCNN models reflects that of our neural processes and, thus, that the representations involved in the former reflect the nature of the neural representations that must carry out those computations in our brains. Indeed, he goes on to argue that understanding neural representation from the standpoint of HCNN allows them to meet the "job description" philosophers of mind require for a physical entity to count as a naturalized representation. Moreover, he goes on to suggest that an insight from HCNN models to understanding intelligent

behavior, is that they reveal not only that there are representations, but also that they are necessary for such a behavior to take place.

Michael Anderson and Heather Champion, both at the University of Western Ontario, disagree with Poldrack. They start their contribution by distinguishing between "mental representations", or the entities posited by the representational theory of mind and invoked by psychological explanations; "artificial representations", or the kinds of entities postulated by artificial network models such as HCNN; and "neural representations", or the entities identified by neuroscience as actually representing their contents. Next, Anderson and Champion go on to characterize Poldrack's project as that of showing how artificial representations (in HCNN) are relevantly similar to neural representations and, thus, that they can fulfill the philosophical requirements stipulated for a representation to count as a naturalized mental representation. However, they argue that Poldrack's account faces an unfortunate dilemma. In order for his argument to work, Poldrack needs that artificial representations are equivalent to neural networks in their capacity to reflect a structural (i.e., physical) isomorphism with relevant features of the world. But this requirement does not obtain, as artificial networks in HCNN exhibit a mathematical, not a physical, isomorphism with the objects they represent, rendering them insufficient to meet the structural isomorphism required for neural representations. However, if Poldrack renounces this requirement, then it looks as though the notion of neural representation is insufficient to fulfill the "job description" a mental representation demands. Thus the dilemma: either artificial representations postulated by HCNN are not equivalent to neural representations, or they don't give us the kinds of representations a representational theory of the mind requires.

Ruth Millikan, a founder of the teleosemantic view discussed at the outset and emeritus faculty at the University of Connecticut, brings her understanding of mental representation to this

contemporary debate over representations in neuroscience. As she sees it, the issue is poorly framed. There is no need to go in search for a new theory of representation to handle recent developments in neuroscience. Teleosemantics can handle these cases. In fact, evidence of its ability to be put to use here as it is elsewhere offers the kind of convergent evidence that further bolsters the view. As used in teleosemantics, Millikan reminds us, "representation" is a functional term. It can be put to use wherever the right sorts of functions are found. This, she argues, is where the work needs to be done. Not in developing a neuroscience-specific sense of representation, but in exploring how these alleged representations are being used and by whom or what. Exploration of the users, consumers, or interpreters of neural representations is key, and as Millikan sees it, points a productive way forward. "The use of representations is an engineering principle, like the use of levers or gears" (Millikan, this issue, p. 2462). The work ahead is in articulating the principles of its application to neuroscience.

Rosa Cao (Stanford) argues in the opposite direction. Rather than taking an established view of mental representation and applying it to neuroscience, as Millikan does, Cao develops a form of *representational pragmatism*, whereby the question of whether there are representations in the brain is determined by the role they play in different research projects in neuroscience. The aim is for an ecumenical view: a wide range of entities and processes can serve as neural representations, provided those representations can be identified and re-identified, and that researchers have the ability to demonstrate how manipulation of these alleged representations brings about functionally-relevant changes in the overall system. Cao's account highlights the important role that investigative tools and techniques, *probes* as she calls them, play in this process. Microelectrodes, fMRI, classifiers and decoders - these are all different tools that neuroscientists use to find causal structure in neural activity. The success of a probe is at least preliminary evidence

of the system's causal structure and the representations involved. Cao articulates a set of constraints critical for assessing a given probe/set of neural representations. These constraints urge careful consideration of which probes we use and why. Are the investigative techniques used well suited to the neural structure being studied? Can the selected probe be used not only to identify underlying structure, but re-identify it? Is there evidence that the structure that can be extracted from the system is actually used by the system? This last question is particularly important and easily overlooked in the increasingly popular use of machine learning techniques to decode neural activity. Cao walks through a set of examples from contemporary neuroscience, illustrating how her representational pragmatism can be put to use.

A traditional view in contemporary philosophy of mind and cognitive science is to think of the brain as an information-processing system. Typically, such a system has been interpreted in computational terms, whereby "computation" is understood in digital terms. This view is so influential that it is embedded in the three-level explanatory framework advocated by Marr (1982). According to this approach, information-processing systems, such as brains, can be analyzed from three levels of description. The top-level is the computational level, which specifies the kind of computational problem(s) the system is trying to solve. The mid-level is the algorithmic level, which specifies the precise algorithm and representational resources it deploys to solve the computational problem(s). And, finally, the bottom-level is that of the implementation, which consists in determining how the actual system –in our case the physical organ of the brain– manages to instantiate such representations and carry out the relevant algorithm. However, as Corey Maley (University of Kansas/Purdue University) reminds us, not all information-processing systems are digital: many of them are analog. But when it comes to analog computation, the three-level framework postulated by Marr breaks down, as the algorithmic/representational level and the

implementational levels collapse into one. And since neurons are likely best seen as analog rather than digital information-processing systems –or so argues Maley– then we need to rethink the relationship between neuronal representation and implementation, not as two independent levels of description, but as a single one.

Favela (Central Florida) argues that the "dynamical hypothesis is undergoing a renaissance in contemporary neuroscience." Appreciating the increasing prevalence of this approach in contemporary neuroscience is critical for mitigating the enthusiasm for representation talk, as the dynamical approach is decidedly non-representational. Favela provides a useful introduction to dynamic systems theory and illustration of the ways it has been put to use in neuroscience throughout its history. In so doing, Favela attempts to make clear that the incorporation of dynamical frameworks is not new; rather, it is becoming more influential. Moreover, his telling of the role of dynamical systems theory in neuroscience highlights its applicability not only for capturing phenomena at the cognitive level, but also for understanding neural mechanisms at a range of lower levels. This is critical to Favela's implicit, background argument that progress in neuroscience often involves moving away from representations. Rather than arguing against their role, Favela simply notes the increase in research frameworks, like motor control, where they are no longer deemed necessary.

In the last couple of decades, multivariate pattern analysis (MVPA) of neuroimaging data have become extremely common, and since their inception, many have heralded them as strategies to identify representations in the brain. In her contribution, Adina Roskies (Dartmouth College) explores a particular version of MVPA, called representational similarity analysis (RSA), whereby patterns of brain activation during experimental tasks are encoded as vectors whose distance or "similarity" from other vectors can be calculated. These distance metrics can be arranged in a

"representational dissimilarity matrix" that can in turn be compared with other similarly produced dissimilarity matrices, coming from other conditions, stimuli, recording devices, populations, or even specimens. The resultant projections in "representational space" can help to identify patterns of neural activity associated with a specific kind of information that remain stable across dissimilarity matrices. Thus, according to Roskies, thinking of semantic or representational contents in terms of high-dimensional representational spaces, as suggested by RSA, offers a powerful strategy to understand mental representations in a naturalistic framework. That does not mean, however, that RSA is the last word when it comes to understanding how the brain represents, for Roskies also identifies some limitations in the semantic interpretations afforded by RSA. Instead, she suggests that RSA should be considered as a powerful tool which, along with other techniques, can get us closer to understanding how brains manage to represent.

A different take on MVPA is offered instead by Bryce Gessell, Benjamin Geib and Felipe De Brigard (Duke University) in their contribution. They start off by reviewing how, for the past two decades, many cognitive neuroscientists have claimed that MVPA of neuroimaging data – particularly fMRI data– allows us to see "what information is represented in a brain region [and] how that information is encoded and organized" (Haxby et al., 2014: 436). Contra this claim, Gessell, Geib and De Brigard offer four philosophical challenges to using MVPA as evidence for neural representation. The first two challenges concern what Sullivan (2010) called a "substantive" notion of representation in neuroscience, whereby the term "representation" plays an explanatory role in virtue of it referring to the bearer of a particular representational content. In essence, these two challenges hark back to concerns about the problem of co-instantiation and misrepresentation (presented in Section 1), as they show how MVPA analyses are inadequate to address them. The other two challenges speak to Sullivan's "weaker" notion of representation in neuroscience,

whereby the term is employed simply to signal co-variation between brain activity and a stimulus. However, even with this minimal sense of representation, Gessell et al argue that MVPA falls short of supporting neural representation for MVPA has difficulties with the orthogonalization of categorical features and the interpretation of null results. As such, they claim that the promise that MVPA can provide evidence for neural representation, is overblown.

Attempts at naturalizing mental representations assume that distinct neural vehicles carry distinct contents, and that the physical interactions between such vehicles implement the functional interactions stipulated by the algorithm. This view, which Daniel Burnston (Tulane University) calls "algorithmic homuncularism", has been the backdrop of many naturalistic theories of mental representation, and has been strongly endorsed by most representational realists, including Shea (2018). In his contribution, though, Burnston argues against algorithmic homuncularism on account that recent neural evidence from neurophysiology shows that neurons exhibit *mixed selectivity*, meaning that they are equally selective to different experimental variables or parameters in the stimuli. To make sense of this phenomenon, neuroscientists make use of analytic strategies –such as principal component analysis and linear discriminant analysis– that do not require particular representational contents to be assigned to spatiotemporally distinct parts of the brain. That does not mean, however, that we should reject representational realism. In fact, Burnston argues that one can deny the idea that representational contents are spatiotemporally distinguishable in the brain while holding true our commitments to the existence of representations and their value in the explanation of behavior.

The final two contributions to our collection address particular kinds of neural representations. Robyn Repko Waller (writing from Iona College, now at University of Sussex), explores neural representations of intentional action. Traditional philosophical debates over the

existence of free will have been rejuvenated by EEG studies that purport to show that the neural correlate of intending to act shows up prior to conscious awareness of the intention to act (Libet 1983, 1985). Waller's paper explores a background issue often overlooked in discussions of this research: should these neural correlates be considered a neural representation of the intention to act? Waller summarizes the philosophical literature on intentional action, building a multi-component functional profile for any such neural realizers; the candidate neural vehicle for intentional action would need to play a role in planning for an action *and* while executing that action. Waller then reviews the neuroscientific evidence and finds it wanting, especially on the latter front. Helpfully, Waller sketches the kinds of studies that could be done to identify and establish more promising accounts of the underlying neural vehicles. In so doing, Waller illustrates the kind of productive convergence that the interdisciplinary debate on these issues can produce.

Jonathan Najenson (Technion) focuses on the engram, the purported neural vehicle for representing memories of past events. Najenson is responding to research in cellular and molecular biology of memory where, with the advent of optogenetic techniques, researchers have made significant advances in identifying and manipulating the mechanisms responsible for individual memories (see Josselyn and Tonegawa 2020 for review). Najenson's work explores how such research can be used to precisify both philosophical and neuroscientific thinking about these mental representations, in terms of both their contents and their vehicles. First, Najenson explores the issue of vehicle localization. Have optogenetic techniques found the engram? Do studies of silent engrams provide a challenge to the standard synaptic account of memory storage? Najenson argues that such interpretation relies on ambiguity in how the term 'accessibility' is understood in the memory literature. Second, Najenson explores whether contemporary engram results offer a specific enough proposal of the representational content of memories to adjudicate between

preservationist and constructivist views of memory content in the philosophy of memory. He argues that the evidence is compatible with both views and that further work must be done, both empirically and theoretically to further advance the debate.

4. Concluding remarks

In the three years that have elapsed since the surge of discussion on this topic took over the Twittersphere, neuroscientists and philosophers have continued to talk in terms of representations in the brain. Unlike the changes in Twitter, however, these have been positive developments. Researchers are increasingly having productive interdisciplinary conversations about whether and how brains can represent. In particular, a number of recent publications in scientific venues have made very clear that lessons from philosophy are helpful in advancing the conversation about the nature of neural representations. For instance, in a recent theoretical piece, Barack and Krakauer (2021) argue that there are broadly two implicit views on the nature of neural computations and representations that are implicit in contemporary work in neuroscience. This theoretical piece, which is very philosophical in nature, already has garnered the attention of many practicing neuroscientists that likely hadn't thought of these foundational issues from a philosophical perspective. Likewise, a recent paper by Baker, Lansdell and Kording (2022) frames contemporary discussions about the nature of representation in neuroscience within the conceptual apparatus, discussed in section 2, which philosophers of mind have developed for understanding naturalized mental representations. Similarly, Piccinini (2022) articulates the philosophical problem of mental content in terms that are translatable to neural representation – and importantly, published this work in *Frontiers in Neurorobotics*, targeting scientists rather that philosophers. These are only a few of the available examples. Any

further delays to the completion of this introduction will only serve to increase the number we lack the space to acknowledge.

It is our hope that the articles included in this topical collection help to move the conversation forward. Neuroscience will continue to develop at a fast pace, with new and more advanced technologies as well as more complex and sophisticated strategies for data analyses. But as the discipline grows it is all the more important to make sure that its foundation is solid and stable. Whether wanted or not, the neuroscientific practice is full of theoretical terms that carry a heavy conceptual baggage. Practitioners may ignore these commitments and even dismiss them as mere "semantics". But it is *precisely* because they are semantic issues–that is, issues that pertain to the things that our neuroscientific concepts refer to– that they cannot be brushed aside. Sooner or later, data needs to be interpreted, and for that concepts are inescapable. Philosophy, a discipline that traditionally has been negotiating with conceptual difficulties, is in an excellent position to help assure that the theoretical foundations of neuroscience are in good shape. The invitation, then, is for neuroscientists and philosophers to work together, as they both have much to learn from each other.

References

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, *33*(4), 245-266.

Aronov, D., Nevers, R., & Tank, D. W. (2017). Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, *543*(7647), 719-722.

Baker, L. R. (1991). Has content been naturalized? In Loewer, B. & Rey, G. *Meaning in Mind: Fodor and his critics.* Blackwell, pp 17-32.

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in cognitive sciences*.

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

Barwich, A. S. (2019). The value of failure in science: The story of grandmother cells in neuroscience. *Frontiers in neuroscience*, *13*, 1121.

Brentano, F. (1874*). Psychology from an Empirical Standpoint*. Routledge.

Brette, R. (2021). Integrative neuroscience of paramecium, a "swimming neuron". *Eneuro*, *8*(3).

Broca, P. (1863). Localisation des fonctions cerebrales: Sie'e du langage articule' *The Bulletin of the Society of Anthropology (Paris)* 4:200–203.

Chisholm, R. M. (1957). Perceiving: A Philosophical Study. Cornell University Press.

Dennett, D. C. (1969). *Content and Consciousness*. Routledge.

Descartes (1664). *Treatise of Man.* Thomas Steele Hall (tr.) (1972). Cambridge: Newcomb Libraria Press.

Dodds, W.J. (1878). On the localization of the function of the brain.

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., ... & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 1-20.

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

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

Finger, S. (2000). *Minds behind the brain: A history of the pioneers and their discoveries*. Oxford University Press.

Fodor, J. A. (1986). Why paramecia don't have mental representations. *Midwest studies in philosophy*, *10*, 3-23.

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

Garson, J. (2015). The birth of information in the brain: Edgar Adrian and the vacuum tube. *Science in Context*, *28*(1), 31-52.

Gates, G. (1996). The price of information. *Synthese*, *107*, 325-347.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature neuroscience*, *3*(2), 191-197.

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801-806.

Haxby, J. V., et al. (2014). A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. Neuron, 72(2), 404-416.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive Fields of Single Neurones in the Cat's Striate Cortex. The Journal of Physiology, 148(3), 574-591.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106.

John, Y. (2021). *In defense of placeholders: the case of 'representation'*. In: https://yohanjohn.com/neurologism/in-defense-of-placeholders-the-case-of-representation/

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of neuroscience*, *17*(11), 4302-4311.

Josselyn, S. A., & Tonegawa, S. (2020). Memory Engrams: Recalling the Past and Imagining the Future. Science, 367(6473), eaaw4325.

Lettvin, Maturana, McCulloch, and Pitts. (1959). What the Frog's Eye Tells the Frog's Brain.

Libet, B. (1983). Time of Conscious Intention to Act in Relation to Onset of Cerebral Activity (Readiness-Potential). The Impending Crisis in the Cognitive Sciences, 23-36.

Libet, B. (1985). Unconscious Cerebral Initiative and the Role of Conscious Will in Voluntary Action. Behavioral and Brain Sciences, 8(4), 529-566.

Loewer, B. (2017). A guide to naturalizing semantics. *A Companion to the Philosophy of Language*, 174-196.

Marr, D. (1982). Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Company.

McCaffrey, J., & Wright, J. (2022). 14 Neuroscience and Cognitive Ontology: A Case for Pluralism.

McCaffrey, J. B. (2015). The brain's heterogeneous functional landscape. *Philosophy of Science*, *82*(5), 1010-1022.

McClelland, J. L., & Jenkins, E. (2014). Nature, nurture, and connections: Implications of connectionist models for cognitive development. In *Architectures for intelligence* (pp. 41-73). Psychology Press.

Millikan, R. G. (1984). Language, Thought, and Other Biological Categories: New Foundations for Realism. The MIT Press.

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

Mundale, J. (2002). Concepts of localization: Balkanization in the brain. *Brain and Mind*, *3*, 313-330.

Neander, K. (1991). The Teleological Notion of Function: A Defense. Pacific Philosophical Quarterly, 72(1), 48-75.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely-moving rat. *Brain research*.

O'Keefe, J & Nadel, L. (1978). *The Hippocampus as a Cognitive Map.* Oxford: Clarendon Press.

Neander, K. (2017). A Mark of the Mental: In Defense of Informational Teleosemantics. The MIT Press.

Paris, J. A. (1831). Pharmacologia, 4th American Edition. *New York, WE Dean*, 79.

Penfield, W., & Perot, P. (1963). The brain's record of auditory and visual experience: a final summary and discussion. *Brain*, *86*(4), 595-696.

Piccinini, G. (2022). Situated neural representations: Solving the problem of content. *Frontiers in Neurorobotics, 16,* doi:10.3389/fnbot.2022.846979.

Pitcher, D., Garrido, L., Walsh, V., & Duchaine, B. C. (2008). Transcranial magnetic stimulation disrupts the perception and embodiment of facial expressions. *Journal of Neuroscience*, *28*(36), 8929-8933.

Quine, W. V. (1960). Word and Object. The MIT Press.

Robins, S. (2023). The 21st century engram. *Wiley Interdisciplinary Reviews: Cognitive Science*, e1653.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, *20*(1), 11.

Shannon, C. E. (1948). A Mathematical Theory of Communication. The Bell System Technical Journal, 27(3), 379-423.

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

Solstad, T., Boccara, C. N., Kropff, E., Moser, M. B., & Moser, E. I. (2008). Representation of geometric borders in the entorhinal cortex. *Science*, *322*(5909), 1865-1868.

Stookey, B. (1993). Jean-Baptiste Bouillaud and Ernest AUBURTIN. Early studies on cerebral localization and the speech center, *Journal of the American Medical Association, 184,* 1024-1029.

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

Uttal, W. R. (2001). *The new phrenology: The limits of localizing cognitive processes in the brain*. The MIT press.

Viola, M. & Zanin, E. (2017). The standard ontological framework of cognitive neuroscience: some lessons from Broca's area. *Philosophical Psychology, 30,* 945-969.

Ward, Z. B. (2023). Muscles or Movements? Representation in the Nascent Brain Sciences. *Journal of the History of Biology*, 1-30.