Philosophical tools to understand conceptual development in neuroscience

Philipp Haueis[1] and Margulies Daniel[2]

[1]Bielefeld University Faculty of History Philosophy and Theology
[2]Centre National de la Recherche Scientifique

April 22, 2024

Abstract

Alongside models and methods, concepts are one of the building blocks of neuroscientific inquiry. They help researchers to pursue various goals, such as describing novel patterns in the data, situate these patterns in existing models, formulate new models or build explanations, e.g., of how the brain processes sensory information. Often the formation of novel or reinterpretation of existing concepts can signal major shifts in how we understand the brain and investigate it scientifically. Yet not every novel concept or reinterpretation produces such a shift, which raises the question of when concepts succeed or fail to change our understanding of the brain. In this paper, we introduce analytical tools developed by philosophers of science and concepts of brain organization as examples to discuss when introducing novel, reinterpreting existing and replacing outdated neuroscientific concepts succeeds or fails. Our discussion opens up novel avenues for neuroscientists and philosophers to collaborate around the limits of old and prospects of new concepts describing how the brain is organized.

Philosophical tools to understand conceptual development in neuroscience

Philipp Haueis[1] †

Daniel S. Margulies[2]

[1] Department of Philosophy, Bielefeld University, P.O. Box 1001331 D- 33501, Bielefeld, Germany

[2] Centre National de la Recherche Scientifique (CNRS) Integrative Neuroscience and Cognition Center, 45, rue des Saints Pères, 75006 Paris, France

† Corresponding author, contact at philipp.haueis@uni-bielefeld.de

ABSTRACT: Alongside models and methods, concepts are one of the building blocks of neuroscientific inquiry. They help researchers to pursue various goals, such as describing novel patterns in the data, situate these patterns in existing models, formulate new models or build explanations, e.g., of how the brain processes sensory information. Often the formation of novel or reinterpretation of existing concepts can signal major shifts in how we understand the brain and investigate it scientifically. Yet not every novel concept or reinterpretation produces such a shift, which raises the question of when concepts succeed or fail to change our understanding of the brain. In this paper, we introduce analytical tools developed by philosophers of science and concepts of brain organization as examples to discuss when introducing novel, reinterpreting existing and replacing outdated neuroscientific concepts succeeds or fails. Our discussion opens up novel avenues for neuroscientists and philosophers to collaborate around the limits of old and prospects of new concepts describing how the brain is organized.

Keywords
Concepts, default mode network, cortical gradient, hierarchy, cortical column

Running title
Conceptual development in neuroscience

Word count: 11280

Figures: 7 (5 color, 2 b/w)

1. Introduction

Alongside models and methods, concepts are one of the building blocks of neuroscientific inquiry. They help researchers to pursue various goals, such as describing novel patterns in the data, situate these patterns in existing models, formulate new models, or build explanations, e.g., of how the brain processes sensory information. Often concepts are tied to specific research questions or are introduced in the wake of developing techniques that enable us to study the brain in novel ways. Thus, when we introduce novel concepts or reinterpret existing ones, our understanding of brain may shift, allowing us to ask to novel questions using specific experimental or modeling techniques. Putative examples such shifts are the development of "feature detector", which allowed electrophysiologists using microelectrode recordings to ask how individual neurons encode sensory information (Barlow 1972, Martin 1994), or the introduction of "default mode network" (Greicius et al. 2003), which was part of a larger shift in functional neuroimaging studies from individual areas towards large-scale cortical networks (Passingham et al. 2002, Biswal et al. 2010, Sporns 2011, Mišić and Sporns 2016).

While conceptual change can certainly shift our understanding of the brain, not every conceptual modification has this effect. Sometimes novel concepts add to a conceptual framework that is already accepted. In other situations, concepts lack flexibility and remain confined to the research context within which they were introduced. Researchers may also propose to reinterpret an existing concept for theoretical reasons but fail to produce evidence which empirically distinguishes this interpretation from alternatives. Finally, concepts can also become outdated over time, such that the understanding of the brain they suggest loses plausibility. So while neuroscientific concepts can be successful by shifting our understanding of the brain, they can also fail to do so in various ways.

In this paper we aim to show that recent philosophical work on conceptual development is helpful to systematically describe and evaluate when concepts succeed or fail to change our understanding of the brain. Philosophers of science have long emphasized that that conceptual

change is a crucial mechanism for scientific progress, but their discussions focused mainly on theory change and cases from physics (Kuhn 1970, Nersessian 1992, 2008). More recently, however, philosophers of neuroscience have argued that change in neuroscience is primarily driven by developing experimental tools rather than theories, and that there are important relations between conceptual development and how tools are developed or used in research (Bickle 2016, 2018, Bickle et al. 2021, Haueis 2023, Colaço and Robins 2023). Since these discussions have mostly appeared in philosophical journals, one of our goals here is to demonstrate the utility of this novel literature to a neuroscientific audience.

Our thesis is that how we should evaluate conceptual development depends on the kind of problem situation that researchers are facing. We systematically distinguish between (1) introducing novel concepts, (2) reinterpreting existing concepts and (3) replacing outdated concepts and formulate conditions under which (1)-(3) succeed or fail to change how we think about and investigate the brain.

To showcase the utility of these conditions, we use concepts from our own neuroscientific and philosophical research, such as "default mode network", "cortical gradient", "hierarchy", or "cortical column" (Callard and Margulies 2011, 2014, Margulies et al. 2016, Hilgetag and Goulas 2020, Bernhardt et al. 2021, Haueis 2021a, b, Burnston and Haueis 2021, Haueis 2023). These concepts exemplify our shared interest in how to conceptualize brain organization, especially in the neocortex, at various spatial scales. Our reflections on these concepts are shaped in part by more than 10 years of interdisciplinary collaboration, which involved visiting laboratory meetings, jointly participating in neuroscientific and philosophical conferences, co-organizing workshops and collaborating in interdisciplinary research projects. Despite our particular interest in concepts of brain organization, we aim to formulate general insights into when conceptual development in neuroscience succeeds or fails.

The paper is organized as follows. Section 2 introduces common features of neuroscientific concepts. In section 3, we discuss when introducing a novel concept succeeds

or fails to initiate change in neuroscience and illustrate our proposal with the concepts "default mode network" and "cortical gradient". Section 4 distinguishes three ways of reinterpreting an existing concept and articulates conditions under which a particular reinterpretation should be favored. The interpretation of "hierarchy" in systems neuroscience will serve as example. In section 5, we consider when a concept should be replaced by an alternative, or even abandoned altogether, and use the example of "cortical column" to illustrate our proposal. We conclude by reflecting on possible collaborations between philosophers and neuroscientists in assessing existing and forming novel concepts that improve our understanding of the brain.

2. Features of neuroscientific concepts

Our discussion draws on three features that philosophers have recently emphasized when analyzing neuroscientific concepts. The first feature is that applying a neuroscientific concept involves a specific experimental or modeling technique. An experimental technique instructs researchers how to use a measurement or intervention device to produce a specific kind of experimental result (Colaço 2018). For example, using "receptive field" in electrophysiological experiments involves extracellular microelectrode recordings to produce spiking patterns in response to sensory stimuli (Hartline 1938, Chirimuuta and Gold 2009). By contrast, using "resting state network" in neuroimaging experiments involves the technique of recording functional activity during the experimental resting state (e.g. with fMRI) and searching for groups of areas in which the recorded signal (e.g. low-frequency BOLD oscillations in fMRI) is correlated over time (Smith et al. 2009). Applying neuroscientific concepts can also involve modeling techniques, which instruct researchers how to use a mathematical equation to produce a specific kind of numerical result. Examples are studies using equations from dynamical systems theory to apply "critical point", or equations from graph theory to apply "centrality" to experimental or simulated brain data (Deco et al. 2013, Deco and Kringelbrach 2017, see Favela

2020, see also section 4.2). Often, different studies use the same concept but apply different techniques, which raises the question how these technique-involving uses are related (see section 3).

The second feature is that neuroscientific concepts target neural properties that are instantiated at different spatial and/or temporal scales (Haueis 2021a). For example, while "receptive field" targets stimulus-specific spike patterns at the microscale of individual neurons, while "default mode network" targets intrinsic brain activity at the macroscale of cortical areas and whole-brain networks. Yet other terms like "cortical column" target properties at the mesoscale of cortical circuits, such as stereotypical connection patterns between different kinds of neurons (Mitra 2014). Over time concepts often get extended from one scale to another. An example are electrophysiologists who extended "cortical column" to target microscale properties of minicolumns and macroscale properties of hypercolumns (Hubel and Wiesel 1974a, Mountcastle 1978). Researchers also aim to link multiple scale-dependent uses of a concept together. One example are studies using "gradient" to link macroscale patterns of functional connectivity to microscale patterns in cytoarchitecture (Huntenburg et al. 2017, Burt et al. 2018, Paquola et al. 2019, Fulcher et al. 2019, Shafei et al. 2020).

The third feature is that neuroscientific concepts can be used to pursue various epistemic goals (Bloch-Mullins 2020). An epistemic goal is a cognitive achievement that a scientific community tries to reach when using a scientific concept (Brigandt 2010). Examples of epistemic goals pursued with the concepts discussed here include: characterizing a newly discovered pattern in the data ("default mode network") situating that pattern within existing models of brain organization ("gradient"), identifying a building block in the neocortex ("cortical column") and explaining sensory information is processed across the brain ("hierarchy"). This supports the general philosophical view that scientific concepts function as tools that can be used for different purposes (Feest 2010, Feest and Steinle 2012, Haueis 2021c). Like material tools, neuroscientific concepts can be fit for one purpose but be unhelpful for

another. This suggests that when evaluating a neuroscientific concept, we should always ask what epistemic goals we aim to reach when using that concept (section 5).

3. Introducing novel concepts in neuroscientific research

3.1 When does introducing concepts in neuroscience succeed or fail?

During neuroscientific discovery, we frequently encounter previously unknown patterns of experimental data or model results. To characterize these patterns adequately, we may choose to introduce a novel term. Yet not every novel concept significantly shifts how we understand the brain. If a concept merely adds to the conceptual framework that the community already accepts, then its potential for change may be limited. Take the example of "feature detector" mentioned above. Electrophysiologists already successfully measured receptive field properties for decades (Shepherd 2010) when Barlow introduced this concept as part of the single-neuron doctrine (Barlow 1972). This suggests that novel concepts can have greater impact if they solve problems or characterize phenomena outside of accepted conceptual frameworks (Arabatzis and Nersessian 2015).

We suggest that concept formation can be particularly impactful during exploratory research. Both neuroscientists and philosophers often characterize exploratory experiments negatively as experiments which do not test a hypothesis derived from a theory (Hubel and Wiesel 1998, Elliot 2007, Haueis 2014, Hussain and Cohen 2017, Colaço 2018).[1] One positive function is that exploratory experiments help researchers form novel concepts because they render accepted conceptual frameworks insufficient to characterize newly discovered phenomena (Steinle 1997, Feest 2012). An example in neuroscience is the formation of the

[1] The use of "hypothesis" in articles does not reliably indicate whether a study is exploratory or confirmatory. Researchers may use this term for presentational purposes even when their study used exploratory research strategies (Rowbottom and Alexander 2012). This use of "hypothesis" usually marks a relation between two features of the system the study investigated (Earman and Salmon 1999), rather than hypotheses derived from a theoretical model of the system under study (Giere et al. 2006). Exploratory experiments in neuroscience are hypothesis-free in the latter but not the former sense.

"default mode network", which emerged outside the framework in which neuroimaging researchers used task-induced increases to characterize the cognitive function of brain areas (section 3.2).

While exploratory experimentation is an important locus of concept formation, not every instance of it successfully changes our understanding of the brain. So what does a concept introduced via exploratory experiments need to achieve to initiate change in neuroscience? Following Haueis (2023) we suggest the concept (a) should be anchored in a property that is significant for pursuing an epistemic goal, and that it (b) should be applicable to novel contexts beyond the one within which it was introduced.[2] Condition (a) is important because we do not simply want any concept, but one which *appropriately* describes the newly found pattern (Steinle 2012, Haueis 2023). Condition (b) is important because in order to initiate change, other researchers need to be able to use the novel concept in subsequent studies.

How do novel concepts formed in exploratory experiments succeed in identifying an anchoring property in the brain? The first step is that we use the right experimental conditions to operationally define the meaning of the concept. Not every experimental condition will do the job equally well. Second, we need to correctly infer the functional significance of this operational definition, e.g., by comparing experimental conditions to real-world conditions. The success of this inference will depend on the experimental conditions used in the first step. Third, we need to use an experimental or modeling technique to fix the reference of the functionally significant operational definition to some scale-dependent property in the brain.

To see how the concrete details of two experiments matter here, compare the introduction of "fly detector" (Barlow 1953) and "bug detector" in the frog's optic nerve (Lettvin et al. 1959). Barlow operationally defined "fly detector" as on-off units using small light spots, whereas Lettvin et al. used convex black discs smaller than 1°. First, light intensity

[2] Focusing on mechanisms, Haueis (2023) restricts (a) to causal properties. Here we expand (a) to capture that neuroscientific concepts can also be anchored in noncausal features such as quantitative measures of topography or brain topology (see section 3.2 and 4.1 for examples).

used by Barlow is not an indispensable parameter to drive optic nerve fiber responses, whereas convexity used by Lettvin et al. is (at least for the relevant fiber subgroup). Second, when frogs catch flies in the real world, they do not track when light disappears, but when a small convex object enters their visual field. Therefore, inferring that convexity detectors are "bug detectors" is correct, whereas inferring that on-off units are "fly detectors" is not. Third, Lettvin et al. used electron microscopy to fix the reference of "bug detector" to small unmyelinated axons in the optic nerve, which Barlow was unaware of because they had not been discovered yet. Hence, "bug detector" successfully identifies an anchoring property whereas "fly detector" does not (Haueis 2023).

When do novel concepts succeed in being open-ended? First, we need to ensure that other researchers can apply the concept beyond the context in which it was formed. One way to do this is to show that the concept is an instance of a more general concept. This happened when electrophysiologists recognized "bug detector" is an instance of the more general concept of a "feature detector" (Barlow 1972, Martin 1994), or when neuroimaging researchers recognized "default mode network" as an instance of "resting state network" (Beckmann et al. 2005, Damoiseaux et al. 2006, Smith et. al. 2009). Second, the concept needs to be adaptable to the empirical details that differ from the context of its introduction. One way a concept is adaptable is when it can be applied under a variety of experimental conditions or involve different experimental or modeling techniques. Consequently, concepts can fail to be adaptable when their operational definition is tied to particular conditions and techniques. For example: the operational definition of the term "default mode" that occurs in "default mode network" fails to be open-ended because it is confined to experiments using wakeful rest and positron emission tomography (Raichle and Snyder 2007).

3.2 Example: how successful was introduction of "default mode network"?

The concept “default mode network” was introduced through a convergence of two investigative pathways in neurophysics (Biswal et al. 1995) and cognitive neuroscience (Shulman et al. 1997). This convergence is described in detail elsewhere (Biswal 2012, Callard and Margulies 2011, Raichle 2015). Greicius et al. (2003) showed that areas whose activity is decreased during goal-directed tasks (cognitive neuroscience pathway) are “functionally connected”, i.e. they exhibit correlated BOLD fluctuations >0.1 Hz (neurophysics pathway). How successful was the introduction of “default mode network” to characterize this discovery?

Using the three steps from section 3.1 shows that using “default mode” in “default mode network” fails to identify an anchoring property (Haueis 2023). “Default mode” is operationally defined as uniform oxygen extraction function, i.e. as an equilibrium between blood flow and the oxygen an area extracts (Raichle et al. 2001). This operational definition, however, fails the first step of identifying anchoring properties because it does not distinguish the eight areas of the Greicius et al. (2003) study from the rest of the brain (Fig. 1, Klein 2014).

INSERT FIGURE 1 HERE

Because of the mismatch, “default mode network” also does not capture the functional significance of these particular areas. The “default mode is characteristic of all brain areas at all times” (Raichle and Snyder 2007). What is characteristic of the Shulman areas is that they deactivate during goal-oriented tasks, which indicates that sustained information processing is present at rest (Gusnard and Raichle 2001). The default mode concept does not capture this sense of functional activity because it describes the brain in terms of metabolism and flow of energy, and not in terms of cognition and information processing. Thus, “default mode network” fails to identify an anchoring property that is significant to describe information processing mechanism(s) in functionally connected areas that deactivate during goal-oriented tasks.

By contrast, “default mode network” succeeds to be open-ended because it identifies a network that can be investigated under a wide variety of experimental conditions, using different experimental techniques. This network can be detected in task conditions related to

different cognitive functions (e.g., mind-wandering, autobiographical memory, internally oriented attention) and many different neurological and mental disorders (e.g., Alzheimer's, autism, anxiety, depression, and schizophrenia, Broyd et al. 2008, Buckner et al. 2008, Smallwood et al. 2021).

The above analysis suggests that the success of introducing "default mode network" was mixed, since Greicius et al.'s introduction of the concept, while open-ended, failed to identify an epistemically significant anchoring property. This leaves us with a conundrum: while our philosophical evaluation suggests a mixed success, neuroscientists widely consider that the discovery of DMN changed our understanding of the brain significantly (Callard and Margulies 2014, Raichle 2015). What explains this discrepancy?

We suggest to explain this discrepancy by arguing that the term "default mode network" is *strategically ambiguous*. On the one hand, "default mode" connects the Greicius study to the cognitive neuroscience pathway in which researchers discovered that these areas are tonically active during the experimental resting state (Raichle et al. 2001). On the other hand, the term is neutral between different cognitive hypotheses, e.g., that these areas support processing tasks related to the "self" (Gusnard and Raichle 2001) or that they are critical for retrieving episodic memories (Greicius et al. 2003). "Default mode network" marks the genuine epistemic uncertainty which of these hypotheses is correct at the time of discovery. The term thus encourages researchers to pursue multiple different hypotheses rather than prematurely closing off alternatives. This suggests that when a study fails to identify anchoring property, choosing a strategically ambiguous term actually increases open-endedness. This value of choosing ambiguous terms to reflect imprecise concepts is known from other disciplines, such as "gene" in molecular biology (Rheinberger 2010, Neto 2020).

While strategic ambiguity explains how the term "default mode network" signals uncertainty, the underlying concept should eventually identify an anchoring property.[3] One proposal comes from Margulies et al. (2016), who introduced the term "gradient" to the neuroimaging community by applying diffusion map embedding, a dimensionality reduction technique to resting state functional connectivity data. The result is an embedding space which positions nodes according to the similarity of their functional connectivity data. The greatest amount of variance, or "principal gradient" runs from primary sensory to DMN regions. According to Margulies et al., this principal gradient reflects a hierarchy of *representational abstraction*. Representational abstraction is defined by distance from sensory input—how directly/indirectly the network connects to sensory systems—and content heterogeneity—how many sensory modalities and cognitive domains the network represents information from. This interpretation allowed Margulies et al. to tentatively identify an anchoring property: because the DMN is furthest from sensory/motor areas, it "processes transmodal information that is unrelated to immediate sensory input" (Margulies et al. 2016), could thereby be characterized by processing at the highest degree of representational abstraction.

The introduction of "cortical gradient" to neuroimaging further supports our analysis of exploratory concept formation in neuroscience. To interpret the principal gradient hierarchically, Margulies et al. first operationally define distance from sensory input by introducing a geodesic distance measure for MRI data. Applying this definition shows that the DMN regions are furthest away from primary sensory areas. The researchers secondly evaluate the functional significance of that definition via a functional meta-analysis which shows that the DMN regions are activated by tasks from various cognitive domains (e.g. autobiographical memory and social cognition). Third, the use of Human Connectome Project and Neurosynth datasets fixes the reference to functional connectivity and areal BOLD activity patterns at the

[3] Philosophers usually distinguish between the term in language and the corresponding concept (as a mental representation) by writing, e.g. "default mode network" and DEFAULT MODE NETWORK. As a result, the epistemic impact of linguistic phenomena like strategic ambiguity can be easily missed.

macroscale of cortical areas and networks.[4] The gradient concept is also open-ended and has been applied using a variety of experimental techniques at multiple scales (Burt et al. 2018, Fulcher et al. 2019, and to model simulated brain data (Demirtas et al. 2019). Finally, the choice of "gradient"—despite the anchoring in representational abstraction—is also strategically ambiguous. On the one hand it signals that new neuroimaging studies using "gradient" on the macroscale are linked to older studies on continuous progressions of anatomical features at the microscale (Sanides 1962). On the other hand, it suggests that neuroimaging research that focuses on individual areas or networks is incomplete—researchers need to take whole-brain trends into account to characterize brain organization adequately. Calling the "principal gradient" the "first axis of functional connectivity variance" would have conceivably hindered the open-ended application of this concept.

Note that the choice of "gradient" is only one option to anchor "default mode network" in an epistemically significant property. It turns on a specific interpretation of other concepts like "hierarchy", which we analyze further in section 4.2. Another option is to subdivide the default mode network in multiple networks which contribute to distinct information processing mechanisms. Buckner and DiNicola (2019) discuss studies in which autobiographical memory tasks lead to increased fMRI activity in parahippocampal cortex, ventral parts of posterior midline and caudal regions of the parietal lobule, whereas theory-of-mind tasks lead to activity in rostral parts of the parietal lobule and the temporoparietal junction. This option does not anchor "default mode network" in one, but multiple properties that are significant to identify cortical information processing mechanisms.

[4] Like with "default mode", the use of neuroimaging may fail to identify actual entities or activities in a mechanism, so multi-scale studies are needed to determine lower-scale referents of "cortical gradient" (Haueis 2021a).

4. Reinterpreting existing concepts in neuroscientific research

4.1. When does reinterpreting a neuroscientific concept succeed or fail?

Besides introducing novel concepts, neuroscience research also advances when researchers develop novel techniques to apply concepts that already exist (Haueis 2021b, Colaço and Robins 2023). But how exactly are the existing and novel technique-involving uses of the same concept related? Does a novel use support an existing interpretation, or does it suggest a reinterpretation? In this section we present three options of relating different uses of a concept, and discuss to what extent they succeed or fail to change our understanding of the brain.

The first option is that a novel use *substantiates* the existing interpretation of a concept. In this case the novel use allows us to generate evidence which adds details to that interpretation. An example is how researchers use optogenetics to apply "memory" (Colaço and Robins 2023). Here the existing interpretation of "memory" is the process of encoding, storing and retrieving information. The use of optogenetics substantiates this interpretation by generating data about previously inaccessible aspects, such as activating a stored memory directly.

The substantiation option has a comparatively low potential to change our understanding of the brain. After all, we believe that the novel use supports what we already assumed about the concept. Substantiating an existing interpretation is thus similar to introducing a concept within an already accepted conceptual framework (section 3.1). Certainly, change is not excluded in this situation. For example, optogenetics researchers have introduced the "silent engram", a stored memory that is only accessible via experimental stimulation (Josselyn and Tonegawa 2020, Najeson 2021). Even if we accept this controversial concept, however, it does not fundamentally challenge the encoding-storage-retrieval interpretation of "memory". Thus, substantiation only provides comparatively low potential for change.

The second option is that a novel use *conflicts* with the existing interpretation of a concept. In this case the novel use produces results that speak against the existing, but not an

alternative interpretation of that concept. Consider using "orientation selectivity" to describe the functional organization of primary visual cortex (V1). Microelectrode studies suggest that V1 orientation selectivity is arranged into parallel slabs (Hubel and Wiesel 1977), while optical imaging studies suggest a pinwheel-like organization (Blasdel 1992). Thus, the two technique-involving uses of "orientation selectivity" conflict with one another, provided that we can rule out confounds as the source of disagreement. A conflict view can be additionally supported by independent evidence, i.e. evidence which is produced by an existing technique under different experimental conditions or model parameters and which that favors the reinterpretation of a concept. For example: showing novel kinds of stimuli to awake animals, electrophysiologists discovered that receptive field sizes in V1 depend on stimulus contrast and history of stimulation (Kapadia et al. 1994, 1999). This evidence favors interpreting "receptive field" as a dynamic entity, rather than as a fixed filter (Chirimuuta and Gold 2009).

Compared to substantiation, conflicting uses have a high potential to change our understanding of the brain. We cannot simultaneously maintain conflicting interpretations of the same concept, but need to decide—at least eventually—which one to adopt. Multiple factors play a role in these decisions: how well the interpretation accommodates available data (Barwich 2018), whether it has potential to solve outstanding conceptual issues (Haueis and Kästner 2022), and whether it can foster further technological and methodological developments (Favela 2022). But *once* we decide that conflicting uses speak in favor of reinterpreting an existing concept, the implications are far reaching. The reinterpretation affects not only our understanding of data and model results in future studies, but also any theory that makes use of the existing interpretation of the concept (see section 4.2).

The third option is that both the existing and the novel interpretation is distinct and legitimate. Such *pluralist* views are warranted when multiple uses contribute to different research goals (goal pluralism), or when they target distinct properties which are significant for pursuing a single goal (property pluralism). An example of goal pluralism is "cortical column",

which can be used to identify a neocortical building block (section 5.1) or to explain hierarchical visual processing (Hubel and Wiesel 1962, Kästner and Haueis 2022). An example of property pluralism is the use of "cortical gradient" to link micro- and macroscale progressions of cortical properties to describe how the brain represents information (Fulcher et al. 2018, Paquola et al. 2019, Haueis 2021a).

Pluralist views have a potential for change that lies between substantiation and conflict. Unlike substantiation, a novel use does not simply result in evidence supporting the existing interpretation of the concept. It either provides evidence that contributes to another research goal, or it provides evidence for the existence a previously unknown property. In either case, the novel use suggests that our existing understanding of the brain was at least incomplete. At the same time, a pluralist relation does not force us to decide which interpretation we should favor. It thus preserves the insights gained under the old interpretation, while adding the advancements of the new one (Chang 2012).

4.2 Example: Reinterpreting "hierarchy" in systems neuroscience

The concept of hierarchy is central to our understanding brain organization, but there are at least two interpretations (Burnston and Haueis 2021). The representational interpretation is the traditional one: it is based on an anatomical hierarchy of feedforward, feedback, and lateral connections which underlies a sequence of input-output relations between brain areas. Techniques-involving uses of "hierarchy" that support this interpretation are electrophysiological recordings of receptive field properties (Hubel and Wiesel 1962) and anatomical tract tracing of layer-specific connections between cortical areas (Felleman and Van Essen 1991). According to the representational interpretation, occupying a specific place in the hierarchy involves representing certain types of information and representing that information for further use elsewhere in the system. Consider the visual hierarchy: V4 is at a higher level than V1 because it represents color categories whereas V1 represents only wavelength. A

different part of V4 represents complex shapes rather than V1's representation of local orientation. This representational interpretation has been extended to other sensory systems, motor systems and association areas (Mesulam 1998, Wessinger et al., 2001, Grafton and Hamilton 2007, Bardre et al. 2010).

The second, newer interpretation is topological and involves mathematical techniques and concepts from graph theory, alongside the aforementioned techniques. An area is at a higher hierarchical level if it is more *central*, i.e. if it has more widespread influence on the network of brain areas. Centrality can be measured differently, e.g. by calculating degree, betweenness centrality, or clustering coefficient, and it underlies graph-theoretical descriptions of hubs and modules (van den Heuvel and Sporns 2013). Crucially, we can apply these measures and definitions to connectivity data without specifying the representational architecture of the brain. For example, applying degree and clustering coefficient to anatomical connectivity data of the macaque visual system reveals an area's influence of to the rest of the network, without specifying "representational stages of streams" (Sporns and da Costa 2005). Similarly, topological centrality can be used to detect modules in "purely data driven way" (Sporns and Betzel 2016). The same is true for dynamic measures. For example, the integration value of a node's activity in response to an event is given by the number of other nodes to which it is functionally connected after the event. Higher integration values thus signal higher influence on the network during the respective time period. One can use this measure to describe hierarchy without specifying the type of information represented (Deco and Kringelbach 2017).

How should we reinterpret "hierarchy" given the distinct technique-involving uses of that concept? The first option is that the topological uses of "hierarchy" substantiate the representational interpretation. Both uses produce results about the same hierarchical organization of the brain. What topological studies add is a more detailed understanding of connectivity, but these support interpreting "hierarchy" as representational and functional specificity. Evidence for substantiation comes from studies showing that graph-theoretical

modularity analyses of anatomical connectivity data align with well-known functional divisions in the brain (Sporns et al. 2007, Hagmann et al. 2008, Meunier et al. 2010). Another example is Margulies et al.'s (2016) use of diffusion embedding to describe gradients of functional connectivity data (section 3.1), which itself is neutral with regard to the representational role of a given node (Haueis and Burnston 2021). The interpretation of the first gradient in terms of representational abstraction substantiates Mesulam's model by situating the DMN at the top of a known representational hierarchy that proceeds from unimodal sensory to transmodal association areas (Fig. 2). If we choose the substantiation view, then we need to only minimally adapt our interpretation of "hierarchy" given recent topological uses of this concept.

INSERT FIGURE 2 HERE

The second option is that topological uses conflict with the representational interpretation because they produce conflicting evidence about the hierarchical positions of brain areas or networks. Consider V4, which the representational interpretation places in the middle of the visual hierarchy (Felleman and Van Essen 1991). By contrast, graph-theoretical analyses of anatomical connectivity show that V4 has one of the highest centrality values in the entire brain (Fig. 3). V4 is thus potentially a more integrative area than supposed by the representational interpretation of "hierarchy". Another example are topological analyses of the posterior anterior gradient in prefrontal cortex, which disconfirm predictions about anatomical connectivity entailed by a representational interpretation of this gradient (Goulas et al. 2014). Furthermore, independent anatomical and physiological evidence speaks against central tenets of the representational interpretation, such as distinct representational stages (Hedge and Van Essen 2007), increasing response times (Capalbo et al. 2008) and stimulus selectivity (Rigotti et al. 2013).

INSERT FIGURE 3 HERE

If we choose the conflict view, then we need to drastically change our understanding of hierarchical brain organization. Instead of talking about representation and abstraction of

information, we should interpret "hierarchy" to connote the influence an area has on the brain, e.g. by broadcasting information more widely (Kötter and Stephan 2003, Deco and Kringelbach 2017). This reinterpretation would have wide-reaching consequences, as it would also affect other theories that recruit the idea of a representational hierarchy (Pylyshyn 2007, Barett 2014).

Finally, different versions of pluralism would hold that representational and topological uses of "hierarchy" are distinct and legitimate. Goal pluralism would claim that the representational use of "hierarchy" contributes to explain how signals are in fact processed by the brain, whereas the topological use helps us understand the efficiency of communication under constraints like minimizing wiring length (van den Heuvel and Sporns 2011). These are recognizably distinct, although related explanatory goals. Property pluralism would claim that the brain in fact instantiates many forms of organization, which are targeted by representational and topological uses of "hierarchy". In contexts where organisms make specific perceptual judgments, hierarchical organization is well described as one of representational specificity. But in contexts of action and deliberation, where evaluative and motivational influences are crucial, more complicated forms of signal processing, that is mediated by topological properties, may be required (Burnston and Haueis 2021).

Both forms of pluralism require an intermediate form of change. The first kind acknowledges that the representational interpretation still contributes to certain explanations, while complementing them with additional ones. The second kind leaves room for the traditional representational picture as one kind of organization that the network can adopt, while insisting that we should stop speaking of it as *the* hierarchy in the brain.

5. Replacing outdated concepts

5.1. When does replacing a neuroscientific concept succeed or fail?

So far we highlighted that neuroscientists introduce novel or interpret existing concepts to pursue specific research goals. Concepts thus function like tools in the process of designing novel experiments, conducting measurements and explaining phenomena (Feest 2010, Feest and Steinle 2012). In ongoing research, however, the goals of inquiry are usually ill-defined and the tools used to pursue them imperfect; both only become into clearer view once a research program develops (Bechtel and Richardson 2010, Feest 2011). It can therefore happen that a scientific concept, even though initially deemed helpful for solving a particular problem, becomes outdated and unhelpful to pursue the goals of the research community. In this situation, we may want to abandon the outdated concept, and replace it with an alternative. Yet history of science teaches us that not every call for replacement succeeds. While biologists successfully replaced "germ plasm" with "gene" to describe the unit of inheritance, calls to replace "gene" with "cistron", "recon" and "muton" failed to catch on (Rheinberger 2010). We now discuss when replacing outdated concepts in neuroscience succeeds or fails, and how it changes our understanding of the brain.

The first question is when a concept is outdated. Like models and theories, concepts are never perfect and are always capable to evolve. We cannot simply point to empirical shortcomings, or incomplete explanations and proclaim a concept is outdated. Future researchers may solve these issues by improving methods or reinterpreting the concept (section 4.1). This suggests that outdated concepts are plagued by problems that persist, even after improved methods and reinterpretation. But even that is not enough: maybe only a single use of the concept is problematic, whereas other uses remain helpful. For example: there is some controversy whether "cortical column" applies to cortical barrels in the rat somatosensory cortex (Catania 2002). But that controversy is local and does not affect uses of "cortical column" targeting other properties (Haueis and Novick 2023). So, an outdated concepts shows

persistent problems across *multiple* technique-involving uses, typically targeting properties at different scales.

We suggest that to successfully replace an outdated concept with an alternative, two conditions need to be fulfilled. First, the outdated concept must permanently fail to contribute to the central epistemic goal for which it was introduced (Haueis 2021b). This condition is important because it demands that the multiple problems of the outdated concept are not disconnected failures. They all hinder us from using it to pursue a particular research goal. But how do we know which goal is the central one, given that we can use the same concept to pursue different goals (Brigandt 2010)? Philosophers have not come up with a principled criterion yet, but a good heuristic is that proponents and critics agree that we should reach this goal when using the concept (see section 5.2 for an example). Secondly, the concept which replaces the outdated concept should avoid the issues that led to the permanent failure of pursuing the central epistemic goal. This condition is important to avoid that our replacement efforts become relabeling exercises that fail to address the underlying issues which made the old concept outdated in the first place.

When both conditions are fulfilled, we can expect that our understanding of the brain changes significantly. It implies that the understanding suggested by the outdated concept is implausible, and provides an alternative understanding suggested by the replacement concept, which now seems more pursuitworthy. Notice, however, that the first condition alone can also successfully initiate change. This happens when we have reason to believe that the central epistemic goal should be abandoned along with the outdated concept. In this case, we come to think the research question itself is ill-posed, and that answering it does not help us better understand how the brain is organized. We now discuss an example in which both these options have been proposed.

5.2 Example: Replacing "cortical column" in electrophysiology

"Cortical column" is a central concept in electrophysiology which refers to vertical structures in the neocortex with similar functional response properties. This concept led to Nobel-prize winning research on the visual system by Hubel and Wiesel (1962, 1977), helped articulate the representational interpretation of "hierarchy" (section 4.2) and led to the view cortical columns are basic building blocks of the mammalian neocortex (Mountcastle 1978, 1997). Yet critics proclaimed that "cortical column" presents an outmoded view of cortical organization and should be replaced (DaCosta and Martin 2010) if not abandoned altogether (Horton and Adams 2005). How successful are these proposals given the conditions introduced above?

INSERT FIGURE 4 HERE

To answer this question, let us first summarize the problems of different uses of the column concept depicted in Fig. 4. The first problem was already known to Mountcastle (1957) when he introduced "cortical column": at the mesoscale of cortical circuits, Nissl-stained sections show no discrete anatomical boundaries that would separate one column from another (Fig. 5a). To solve this issue, Mountcastle (1978) introduced "minicolumn" and proposed that vertical cell bands at the microscale are separated because their vertical connections are stronger than their horizontal connections. The proposal was initially empirically supported (Peters and Sethares 1996). But improved methods showed that the problem of missing boundaries persisted because 90% of synapses connect with neurons further than 100μm away from the minicolumn (Fig. 5 right, see also Kaas 2012).

INSERT FIG 5 HERE

The second problem is that while in some areas, functional responses to stimuli are organized in a columnar fashion (e.g. primary somatosensory cortex and V1 in cats and macaques), non-columnar responses are present in many areas. In primary auditory cortex, multiple feature maps exist but they are often layer-specific (Fig. 6). Thus, there seems no evidence for a uniformly sized auditory column (Swindale, 1990). In visual area MT, direction-

selectivity does not change discretely and is arranged in a variable fashion. Additionally, disparity selective cells are distributed in a patchy non-columnar manner over the direction-selectivity map (Fig. 6). This problem of persists despite improved methods such as cytochrome oxidase staining (Horton and Adams 2005).

INSERT FIGURE 6 HERE

The third problem is widespread inter- and intraspecies variation of columnar structures (Barbas et al. 2022). Orientation columns are found in many mammals but are absent in rats and squirrels, even though these animals have cells that are highly orientation-selective (Horton & Adams 2005). Racoons, beavers and cats have whiskers but lack barrels, while guinea pigs have barrels but do not show whisking behavior (Horton and Adams 2005). Similarly, ocular dominance columns are present in some primates, but not others, while both have similar visual capacities (Purves et al, 1992). In squirrel monkeys, 30% of the cases examined lack ocular dominance structures (Horton and Adams 2005). In other individuals of the same species, only part of V1 exhibited ocular dominance columns, despite intact vision.

Taken together, the three problems suggest that cortical column is an outdated concept which permanently fails to contribute to the central epistemic goal for which it was introduced. Both critics and proponents of the concept agree that "cortical column" should identify an anatomically discrete, functionally modular building block that executes the same computation across areas and species (Mountcastle 1978, 1997, Swindale 1990, Purves et al, 1992, Catania 2002, Horton and Adams 2005). This epistemic goal shows that the three problems are not disconnected failures. The problem of missing boundaries questions that we can distinguish neocortical building blocks anatomically (da Costa and Martin 2010). The problem of non-columnar responses questions whether columnarity is the only form of functional organization in neocortex (Catania 2002). The problem of species variation questions that columnar structure is universally present across mammalian species (Horton and Adams 2005). Together, the three

problems also make it highly unlikely that columnar structures are species-invariant units that compute the same function across varying inputs.

The column concept fulfills the first condition for replacement because its failures persist despite improved methods and cannot be resolved by reinterpretation. Consider that we reinterpret "cortical column" in a pluralistic manner, according to which different uses refer to diverse kinds of cortical structure, some of which contribute to information processing whereas others may not (Haueis 2021b). On this 'diverse-kinds' interpretation, however, "cortical column" still fails to identify a neocortical building block. Rather, the column concept refers to domain-specific kinds of structures in distinct species at different scales. These structures may instantiate a property targeted by one use of "cortical column" without instantiating the other. We can thus accept the diverse-kinds interpretation without assuming that "cortical column" refers to a neocortical building block.

Turning to the second condition, da Costa and Martin (2010) suggest to replace "cortical column" with "canonical microcircuit". This concept can be used to identify a building block because it describes neocortical circuit organization in terms of stereotypical connection patterns between excitatory and inhibitory neuron types (Fig. 7). According to its proponents, "canonical microcircuit" contributes to this goal while avoiding the first and second problem that plague the column concept. First, "canonical microcircuit" avoids the problem of missing boundaries by representing circuit structure without specifying a discrete spatial boundary at the mesoscale, which does not exist in neocortex (Fig. 5a). Second, the concept avoids the problem of non-columnar responses by describing cells with functionally heterogenous response properties which do not all connect in a "like-to-like" manner (da Costa and Martin 2010). Regarding the third problem, while some comparisons of circuit organization between cat, macaque and rodent neocortex exists (Beul and Hilgetag 2015), it is currently an open question whether "canonical microcircuit" describes circuit organization that is universal across mammals.

INSERT FIGURE 7 HERE

Depending on which of these conditions we accept, we should change our understanding of neocortical circuit organization more or less drastically. If we reject both conditions, then we can continue to claim that "cortical column" identifies a neocortical building block. To do so, we would need to find ways to overcome three problems outlined above. If we accept the first condition but reject the second, we can adopt the diverse-kinds interpretation of "cortical column", but have to give up the goal of finding a neocortical building block at the mesoscale. Under this interpretation, "cortical column" remains helpful for other goals in current research, such as validating new experimental techniques (Yacoub et al. 2008, Pizutti et al. 2023) or studying evolution and development of columnar structures (Kaas 2012, Barbas et al. 2022). But when it comes to describing circuit organization, we should stop using it and should rather build area- and species-specific descriptions of cortical circuits (Horton and Adams 2005, Haueis 2022b). If we accept both conditions, then we should replace "cortical column" with "canonical microcircuit" and understand the neocortical building block as a non-discrete infrastructure whose elements are dynamically recruited at any given moment (da Costa and Martin 2010).

6. Conclusion

In this paper, we have argued that the analytic toolkit from philosophy of science is helpful to systematically describe and evaluate when concepts succeed or fail to change our understanding of the brain. In particular, we showed that when we evaluate neuroscientific concepts, we need to distinguish between the problem situation of introducing a novel concept from the reinterpretation of existing concepts, and the question when to abandon or replace outdated concepts. We specified conditions under which introducing, reinterpreting or replacing concepts succeeds or fails, and have illustrated the utility of these conditions with examples of concepts

of brain organization at the micro- meso- and macroscale. While many of our examples attest that experimental tools play an important role in conceptual development in the neurosciences (Bickle 2018), we wish to emphasize that mathematical techniques to analyze and model brain data are equally important (Favela 2022). The success conditions put forward in this paper are thus equally applicable to tool-driven and theory-driven conceptual development in the neurosciences.

The framework proposed in this paper also suggests that assessing existing and forming novel concepts for understanding the brain can be a fruitful locus of interdisciplinary collaboration between neuroscience and philosophy. While concepts at the interface between both disciplines like "representation", "computation", "memory" or "emotion" are already scrutinized from both sides, we have shown in this paper that concepts of brain organization can also be fruitfully analyzed from an interdisciplinary perspective. Intensifying the debate around these concepts can thus help improve existing, and potentially even create novel conceptualizations of how the brain is organized.

7. References

Arabatzis, T., & Nersessian, N. J. (2015). Concepts Out of Theoretical Contexts. In T. Arabatzis, J. Renn, & A. Simões (Eds.), *Relocating the History of Science* (Vol. 312, pp. 225–238). Springer International Publishing. https://doi.org/10.1007/978-3-319-14553-2_15

Badre, D., Kayser, A. S., & D'Esposito, M. (2010). Frontal Cortex and the Discovery of Abstract Action Rules. *Neuron*, *66*(2), 315–326. https://doi.org/10.1016/j.neuron.2010.03.025

Barbas, H., Zikopoulos, B., & John, Y. J. (2022). The inevitable inequality of cortical columns. *Frontiers in Systems Neuroscience*, *16*, 921468. https://doi.org/10.3389/fnsys.2022.921468

Barlow, H. B. (1953). Action potentials from the frog's retina. *The Journal of Physiology*, *119*(1), 58–68. https://doi.org/10.1113/jphysiol.1953.sp004828

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perceptual psychology? *Perception*, *1*(4), 371–394. https://doi.org/10.1068/p010371

Barrett, H. C. (2015). *The shape of thought: How mental adaptations evolve*. Oxford university press.

Barwich, A.-S. (2018). How to be rational about empirical success in ongoing science: The case of the quantum nose and its critics. *Studies in History and Philosophy of Science Part A*, *69*, 40–51. https://doi.org/10.1016/j.shpsa.2018.02.005

Barwich, A.-S. (2019). The Value of Failure in Science: The Story of Grandmother Cells in Neuroscience. *Frontiers in Neuroscience*, *13*. https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2019.01121

Bechtel, W., & Richardson, R. C. (2010). *Discovering Complexity: Decomposition and Localization as Strategies in Scientific Research*. The MIT Press. https://doi.org/10.7551/mitpress/8328.001.0001

Beckmann, C. F., DeLuca, M., Devlin, J. T., & Smith, S. M. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1457), 1001–1013. https://doi.org/10.1098/rstb.2005.1634

Beul, S. F., & Hilgetag, C. C. (2015). Towards a â€œcanonicalâ€ agranular cortical microcircuit. *Frontiers in Neuroanatomy*, *8*. https://doi.org/10.3389/fnana.2014.00165

Bickle, J. (2016). Revolutions in Neuroscience: Tool Development. *Frontiers in Systems Neuroscience*, *10*. https://www.frontiersin.org/article/10.3389/fnsys.2016.00024
Bickle, J. (2018). From Microscopes to Optogenetics: Ian Hacking Vindicated. *Philosophy of Science*, *85*(5), 1065–1077. https://doi.org/10.1086/699760
Biswal, B., Yetkin, F. Z., Haughton, V. M., & Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. *Magnetic Resonance in Medicine*, *34*(4), 537–541. https://doi.org/10.1002/mrm.1910340409
Blasdel, G. G. (1992). Differential imaging of ocular dominance and orientation selectivity in monkey striate cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *12*(8), 3115–3138. https://doi.org/10.1523/JNEUROSCI.12-08-03115.1992
Bloch-Mullins, C. L. (2020). Scientific Concepts as Forward-Looking: How Taxonomic Structure Facilitates Conceptual Development. *Journal of the Philosophy of History*, *14*(2), 205–231. https://doi.org/10.1163/18722636-12341438
Brigandt, I. (2010). The epistemic goal of a concept: Accounting for the rationality of semantic change and variation. *Synthese*, *177*(1), 19–40. https://doi.org/10.1007/s11229-009-9623-8
Broyd, S. J., Demanuele, C., Debener, S., Helps, S. K., James, C. J., & Sonuga-Barke, E. J. S. (2009). Default-mode brain dysfunction in mental disorders: A systematic review. *Neuroscience & Biobehavioral Reviews*, *33*(3), 279–296. https://doi.org/10.1016/j.neubiorev.2008.09.002
Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The Brain's Default Network. *Annals of the New York Academy of Sciences*, *1124*(1), 1–38. https://doi.org/10.1196/annals.1440.011
Buckner, R. L., & DiNicola, L. M. (2019). The brain's default network: Updated anatomy, physiology and evolving insights. *Nature Reviews Neuroscience*, *20*(10), Article 10. https://doi.org/10.1038/s41583-019-0212-7
Burnston, D. C., & Haueis, P. (2021). Evolving Concepts of "Hierarchy" in Systems Neuroscience. In F. Calzavarini & M. Viola (Eds.), *Neural Mechanisms* (Vol. 17, pp. 113–141). Springer International Publishing. https://doi.org/10.1007/978-3-030-54092-0_6
Burt, J. B., Demirtaş, M., Eckner, W. J., Navejar, N. M., Ji, J. L., Martin, W. J., Bernacchia, A., Anticevic, A., & Murray, J. D. (2018). Hierarchy of transcriptomic specialization across human cortex captured by structural neuroimaging topography. *Nature Neuroscience*, *21*(9), Article 9. https://doi.org/10.1038/s41593-018-0195-0
Callard, F., & Margulies, D. S. (2011). The Subject at Rest: Novel conceptualizations of self and brain from cognitive neuroscience's study of the 'resting state.' *Subjectivity*, *4*(3), 227–257. https://doi.org/10.1057/sub.2011.11
Callard, F., & Margulies, D. S. (2014). What we talk about when we talk about the default mode network. *Frontiers in Human Neuroscience*, *8*. https://www.frontiersin.org/articles/10.3389/fnhum.2014.00619
Chang, H. (2012). *Is Water H2O?: Evidence, Realism and Pluralism* (Vol. 293). Springer Netherlands. https://doi.org/10.1007/978-94-007-3932-1
Chirimuuta, M., & Gold, I. (2009). The Embedded Neuron, the Enactive Field? In J. Bickle (Ed.), *The Oxford Handbook of Philosophy and Neuroscience* (1st ed., pp. 200–225). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195304787.003.0010
Colaço, D. (2018). Rethinking the role of theory in exploratory experimentation. *Biology & Philosophy*, *33*(5), 38. https://doi.org/10.1007/s10539-018-9648-9
Colaço, D., & Robins, S. (2023). Why have "revolutionary" tools found purchase in memory science? *Philosophy and the Mind Sciences*, *4*. https://doi.org/10.33735/phimisci.2023.10499
Da Costa, N. M. (2010). Whose cortical column would that be? *Frontiers in Neuroanatomy*. https://doi.org/10.3389/fnana.2010.00016
da F. Costa, L., & Sporns, O. (2005). Hierarchical features of large-scale cortical connectivity. *The European Physical Journal B - Condensed Matter and Complex Systems*, *48*(4), 567–573. https://doi.org/10.1140/epjb/e2006-00017-1
Damoiseaux, J. S., Rombouts, S. A. R. B., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., & Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects. *Proceedings of the National Academy of Sciences*, *103*(37), 13848–13853. https://doi.org/10.1073/pnas.0601417103
DeAngelis, G. C., & Newsome, W. T. (1999). Organization of Disparity-Selective Neurons in Macaque Area MT. *The Journal of Neuroscience*, *19*(4), 1398–1415. https://doi.org/10.1523/JNEUROSCI.19-04-01398.1999
Deco, G., & Kringelbach, M. L. (2017). Hierarchy of Information Processing in the Brain: A Novel "Intrinsic Ignition" Framework. *Neuron*, *94*(5), 961–968. https://doi.org/10.1016/j.neuron.2017.03.028
Elliott, K. C. (2007). Varieties of Exploratory Experimentation in Nanotoxicology. *History and Philosophy of the Life Sciences*, *29*(3), 313–336. https://www.jstor.org/stable/23334264
Favela, L. H. (2020). Dynamical systems theory in cognitive science and neuroscience. *Philosophy Compass*, *15*(8), e12695. https://doi.org/10.1111/phc3.12695
Favela, L. H. (2022). "It takes two to make a thing go right": The coevolution of technological and mathematical tools in neuroscience. In J. Bickle, C. F. Craver, & A.-S. Barwich (Eds.), *The tools of neuroscience experiment: Philosophical and scientific perspectives*. Routledge, Taylor & Francis. https://doi.org/10.4324/9781003251392
Feest, U. (2011). Remembering (Short-Term) Memory: Oscillations of an Epistemic Thing. *Erkenntnis*, *75*(3), 391–411. https://doi.org/10.1007/s10670-011-9341-8

Feest, U. (2012). Exploratory Experiments, Concept Formation, and Theory Construction in Psychology. In *Exploratory Experiments, Concept Formation, and Theory Construction in Psychology* (pp. 167–190). De Gruyter. https://doi.org/10.1515/9783110253610.167
Felleman, D. J., & Van Essen, D. C. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, *1*(1), 1–47. https://doi.org/10.1093/cercor/1.1.1-a
Giere, R. N. (Ed.). (1992). *Cognitive models of science*. Univ. of Minnesota Press.
Giere, R. N., Bickle, J., & Mauldin, R. (2006). *Understanding scientific reasoning* (5th ed). Thomson/Wadsworth.
Goulas, A., Uylings, H. B. M., & Stiers, P. (2014). Mapping the hierarchical layout of the structural network of the macaque prefrontal cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, *24*(5), 1178–1194. https://doi.org/10.1093/cercor/bhs399
Grafton, S. T., & De C. Hamilton, A. F. (2007). Evidence for a distributed hierarchy of action representation in the brain. *Human Movement Science*, *26*(4), 590–616. https://doi.org/10.1016/j.humov.2007.05.009
Greicius, M. D., Krasnow, B., Reiss, A. L., & Menon, V. (2003). Functional connectivity in the resting brain: A network analysis of the default mode hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(1), 253–258. https://doi.org/10.1073/pnas.0135058100
Gusnard, D. A., & Raichle, M. E. (2001). Searching for a baseline: Functional imaging and the resting human brain. *Nature Reviews Neuroscience*, *2*(10), Article 10. https://doi.org/10.1038/35094500
Hagmann, P., Cammoun, L., Gigandet, X., Meuli, R., Honey, C. J., Wedeen, V. J., & Sporns, O. (2008). Mapping the structural core of human cerebral cortex. *PLoS Biology*, *6*(7), e159. https://doi.org/10.1371/journal.pbio.0060159
Haueis, P. (2014). Meeting the brain on its own terms. *Frontiers in Human Neuroscience*, *8*. https://www.frontiersin.org/article/10.3389/fnhum.2014.00815
Haueis, P. (2021a). Multiscale modeling of cortical gradients: The role of mesoscale circuits for linking macro- and microscale gradients of cortical organization and hierarchical information processing. *NeuroImage*, *232*, 117846. https://doi.org/10.1016/j.neuroimage.2021.117846
Haueis, P. (2021b). The death of the cortical column? Patchwork structure and conceptual retirement in neuroscientific practice. *Studies in History and Philosophy of Science Part A*, *85*, 101–113.
https://doi.org/10.1016/j.shpsa.2020.09.010
Haueis, P. (2021c). A generalized patchwork approach to scientific concepts. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/716179
Haueis, P. (2023). Exploratory Concept Formation and Tool Development in Neuroscience. *Philosophy of Science*, *90*(2), 354–375. https://doi.org/10.1017/psa.2022.79
Haueis, P., & Kästner, L. (2022). Mechanistic inquiry and scientific pursuit: The case of visual processing. *Studies in History and Philosophy of Science*, *93*, 123–135. https://doi.org/10.1016/j.shpsa.2022.03.007
Hegdé, J., & Van Essen, D. C. (2007). A Comparative Study of Shape Representation in Macaque Visual Areas V2 and V4. *Cerebral Cortex*, *17*(5), 1100–1116. https://doi.org/10.1093/cercor/bhl020
Hilgetag, C. C., & Goulas, A. (2020). 'Hierarchy' in the organization of brain networks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1796), 20190319. https://doi.org/10.1098/rstb.2019.0319
Horton, J. C., & Adams, D. L. (2005). The cortical column: A structure without a function. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 837–862. https://doi.org/10.1098/rstb.2005.1623
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), 106–154. https://doi.org/10.1113/jphysiol.1962.sp006837
Hubel, D. H., & Wiesel, T. N. (1974). Sequence regularity and geometry of orientation columns in the monkey striate cortex. *The Journal of Comparative Neurology*, *158*(3), 267–293. https://doi.org/10.1002/cne.901580304
Hubel, D. H., & Wiesel, T. N. (1977). Ferrier lecture—Functional architecture of macaque monkey visual cortex. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *198*(1130), 1–59. https://doi.org/10.1098/rspb.1977.0085
Hubel, D. H., & Wiesel, T. N. (1998). Early Exploration of the Visual Cortex. *Neuron*, *20*(3), 401–412. https://doi.org/10.1016/S0896-6273(00)80984-8
Huntenburg, J. M., Bazin, P.-L., Goulas, A., Tardif, C. L., Villringer, A., & Margulies, D. S. (2017). A Systematic Relationship Between Functional Connectivity and Intracortical Myelin in the Human Cerebral Cortex. *Cerebral Cortex*, *27*(2), 981–997. https://doi.org/10.1093/cercor/bhx030
Hussain, S. J., & Cohen, L. G. (2017). Exploratory studies: A crucial step towards better hypothesis-driven confirmatory research in brain stimulation. *The Journal of Physiology*, *595*(4), 1013–1014. https://doi.org/10.1113/JP273582
Josselyn, S. A., & Tonegawa, S. (2020). Memory engrams: Recalling the past and imagining the future. *Science*. https://doi.org/10.1126/science.aaw4325
Kaas, J. H. (2012). Evolution of columns, modules, and domains in the neocortex of primates. *Proceedings of the National Academy of Sciences*, *109*(supplement_1), 10655–10660. https://doi.org/10.1073/pnas.1201892109
Kapadia, M., Gilbert, C., & Westheimer, G. (1994). A quantitative measure for short-term cortical plasticity in human vision. *The Journal of Neuroscience*, *14*(1), 451–457. https://doi.org/10.1523/JNEUROSCI.14-01-00451.1994

Kapadia, M. K., Westheimer, G., & Gilbert, C. D. (1999). Dynamics of spatial summation in primary visual cortex of alert monkeys. *Proceedings of the National Academy of Sciences*, *96*(21), 12073–12078. https://doi.org/10.1073/pnas.96.21.12073
Klein, C. (2014). The Brain at Rest: What It Is Doing and Why That Matters. *Philosophy of Science*, *81*(5), 974–985. https://doi.org/10.1086/677692
Kötter, R., & Stephan, K. E. (2003). Network participation indices: Characterizing component roles for information processing in neural networks. *Neural Networks: The Official Journal of the International Neural Network Society*, *16*(9), 1261–1275. https://doi.org/10.1016/j.neunet.2003.06.002
Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). University of Chicago Press.
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 IRE*, *47*(11), 1940–1951. https://doi.org/10.1109/JRPROC.1959.287207
Linden, J. F., & Schreiner, C. E. (2003). Columnar Transformations in Auditory Cortex? A Comparison to Visual and Somatosensory Cortices. *Cerebral Cortex*, *13*(1), 83–89. https://doi.org/10.1093/cercor/13.1.83
Margulies, D. S., Ghosh, S. S., Goulas, A., Falkiewicz, M., Huntenburg, J. M., Langs, G., Bezgin, G., Eickhoff, S. B., Castellanos, F. X., Petrides, M., Jefferies, E., & Smallwood, J. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences*, *113*(44), 12574–12579. https://doi.org/10.1073/pnas.1608282113
Mars, R. B., Passingham, R. E., & Jbabdi, S. (2018). Connectivity Fingerprints: From Areal Descriptions to Abstract Spaces. *Trends in Cognitive Sciences*, *22*(11), 1026–1037. https://doi.org/10.1016/j.tics.2018.08.009
Martin, K. A. (1994). A brief history of the "feature detector." *Cerebral Cortex (New York, N.Y.: 1991)*, *4*(1), 1–7. https://doi.org/10.1093/cercor/4.1.1
Mesulam, M. M. (1998). From sensation to cognition. *Brain: A Journal of Neurology*, *121 (Pt 6)*, 1013–1052. https://doi.org/10.1093/brain/121.6.1013
Meunier, D., Lambiotte, R., & Bullmore, E. T. (2009). Hierarchical modularity in human brain functional networks. *Frontiers in Neuroinformatics*, *3*. https://doi.org/10.3389/neuro.11.037.2009
Mišić, B., & Sporns, O. (2016). From regions to connections and networks: New bridges between brain and behavior. *Current Opinion in Neurobiology*, *40*, 1–7. https://doi.org/10.1016/j.conb.2016.05.003
Mountcastle, V. (1997). The columnar organization of the neocortex. *Brain*, *120*(4), 701–722. https://doi.org/10.1093/brain/120.4.701
Mountcastle, V. B. (1978). An organizing principle for cerebral function: The unit module and the distributed system. In G. M. Edelman & V. B. Mountcastle, *The mindful brain: Cortical organization and the group-selective theory of higher brain function* (pp. 7–50). MIT Press.
Najenson, J. (2021). What have we learned about the engram? *Synthese*, *199*(3), 9581–9601. https://doi.org/10.1007/s11229-021-03216-2
Nersessian, N. J. (1992). How do Scientists Think? Capturing the Dynamics of Conceptual Change in Science. In R. N. Giere (Ed.), *Cognitive models of science* (pp. 3–45). Univ. of Minnesota Press.
Neto, C. (2020). When imprecision is a good thing, or how imprecise concepts facilitate integration in biology. *Biology & Philosophy*, *35*(6), 58. https://doi.org/10.1007/s10539-020-09774-y
Novick, R., & Haueis, P. (2023). Patchworks and operations. *European Journal for Philosophy of Science*, *13*(1), 15. https://doi.org/10.1007/s13194-023-00515-y
Paquola, C., Vos De Wael, R., Wagstyl, K., Bethlehem, R. A. I., Hong, S.-J., Seidlitz, J., Bullmore, E. T., Evans, A. C., Misic, B., Margulies, D. S., Smallwood, J., & Bernhardt, B. C. (2019). Microstructural and functional gradients are increasingly dissociated in transmodal cortices. *PLOS Biology*, *17*(5), e3000284. https://doi.org/10.1371/journal.pbio.3000284
Passingham, R. E., Stephan, K. E., & Kötter, R. (2002). The anatomical basis of functional localization in the cortex. *Nature Reviews Neuroscience*, *3*(8),. https://doi.org/10.1038/nrn893
Pizzuti, A., Huber, L. (Renzo), Gulban, O. F., Benitez-Andonegui, A., Peters, J., & Goebel, R. (2023). Imaging the columnar functional organization of human area MT+ to axis-of-motion stimuli using VASO at 7 Tesla. *Cerebral Cortex*, *33*(13), 8693–8711. https://doi.org/10.1093/cercor/bhad151
Powell, T. P., & Mountcastle, V. B. (1959). Some aspects of the functional organization of the cortex of the postcentral gyrus of the monkey: A correlation of findings obtained in a single unit analysis with cytoarchitecture. *Bulletin of the Johns Hopkins Hospital*, *105*, 133–162.
Purves, D., Riddle, D. R., & LaMantia, A.-S. (1992). Iterated patterns of brain circuitry (or how the cortex gets its spots). *Trends in Neurosciences*, *15*(10), 362–368. https://doi.org/10.1016/0166-2236(92)90180-G
Pylyshyn, Z. W. (2007). *Things and Places: How the Mind Connects with the World*. The MIT Press. https://doi.org/10.7551/mitpress/7475.001.0001
Raichle, M. E. (2015). The Brain's Default Mode Network. *Annual Review of Neuroscience*, *38*(1), 433–447. https://doi.org/10.1146/annurev-neuro-071013-014030
Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2), 676–682. https://doi.org/10.1073/pnas.98.2.676

Raichle, M. E., & Snyder, A. Z. (2007). A default mode of brain function: A brief history of an evolving idea. *NeuroImage*, *37*(4), 1083–1090. https://doi.org/10.1016/j.neuroimage.2007.02.041
Rheinberger, H.-J. (2010). The Concept of the Gene. Molecular Biological Perspectives. In H.-J. Rheinberger, *An Epistemology of the Concrete: Twentieth-Century Histories of Life*. Duke University Press. https://doi.org/10.1515/9780822391333
Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160
Robins, S. K. (2018). Memory and Optogenetic Intervention: Separating the Engram from the Ecphory. *Philosophy of Science*, *85*(5), 1078–1089. https://doi.org/10.1086/699692
Rowbottom, D. P., & Alexander, R. M. (2012). The Role of Hypotheses in Biomechanical Research. *Science in Context*, *25*(2), 247–262. https://doi.org/10.1017/S0269889712000051
Salmon, W. C., & Earman, J. (1999). The confirmation of scientific hypotheses. In M. H. Salmon, J. Earman, C. Glymour, J. Lennox, P. Machamer, J. E. McGuire, J. D. Norton, W. C. Salmon, & K. C. Schaffner, *Introduction to the philosophy of science* (Reprinted). Hackett Publ.
Sanides, F. (1962). *Die Architektonik des Menschlichen Stirnhirns.* Springer.
Shafiei, G., Markello, R. D., Vos de Wael, R., Bernhardt, B. C., Fulcher, B. D., & Misic, B. (2020). Topographic gradients of intrinsic dynamics across neocortex. *eLife*, *9*, e62116. https://doi.org/10.7554/eLife.62116
Shepherd, G. M. (2010). *Creating modern neuroscience: The revolutionary 1950s*. Oxford University Press.
Shulman, G. L., Fiez, J. A., Corbetta, M., Buckner, R. L., Miezin, F. M., Raichle, M. E., & Petersen, S. E. (1997). Common Blood Flow Changes across Visual Tasks: II. Decreases in Cerebral Cortex. *Journal of Cognitive Neuroscience*, *9*(5), 648–663. https://doi.org/10.1162/jocn.1997.9.5.648
Smallwood, J., Bernhardt, B. C., Leech, R., Bzdok, D., Jefferies, E., & Margulies, D. S. (2021). The default mode network in cognition: A topographical perspective. *Nature Reviews Neuroscience*, *22*(8), Article 8. https://doi.org/10.1038/s41583-021-00474-4
Sporns, O. (2011). *Networks of the brain*. Massachusetts institute of technology.
Sporns, O., & Betzel, R. F. (2016). Modular Brain Networks. *Annual Review of Psychology*, *67*(1), 613–640. https://doi.org/10.1146/annurev-psych-122414-033634
Sporns, O., Honey, C. J., & Kötter, R. (2007). Identification and Classification of Hubs in Brain Networks. *PLoS ONE*, *2*(10), e1049. https://doi.org/10.1371/journal.pone.0001049
Steinle, F. (2012). Goals and Fates of Concepts: The Case of Magnetic Poles. In *Scientific Concepts and Investigative Practice* (pp. 105–126). De Gruyter. https://doi.org/10.1515/9783110253610.105
Steinle, F., & Feest, U. (Eds.). (2012). *Scientific Concepts and Investigative Practice: Introduction*. De Gruyter. https://doi.org/10.1515/9783110253610
Swindale, N. V. (1990). Is the cerebral cortex modular? *Trends in Neurosciences*, *13*(12), 487–492. https://doi.org/10.1016/0166-2236(90)90082-L
Van Den Heuvel, M. P., & Sporns, O. (2011). Rich-Club Organization of the Human Connectome. *The Journal of Neuroscience*, *31*(44), 15775–15786. https://doi.org/10.1523/JNEUROSCI.3539-11.2011
van den Heuvel, M. P., & Sporns, O. (2013). Network hubs in the human brain. *Trends in Cognitive Sciences*, *17*(12), 683–696. https://doi.org/10.1016/j.tics.2013.09.012
Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, *13*(1), 1–7. https://doi.org/10.1162/089892901564108
Yacoub, E., Harel, N., & Uğurbil, K. (2008). High-field fMRI unveils orientation columns in humans. *Proceedings of the National Academy of Sciences*, *105*(30), 10607–10612. https://doi.org/10.1073/pnas.0804110105

8. Figures

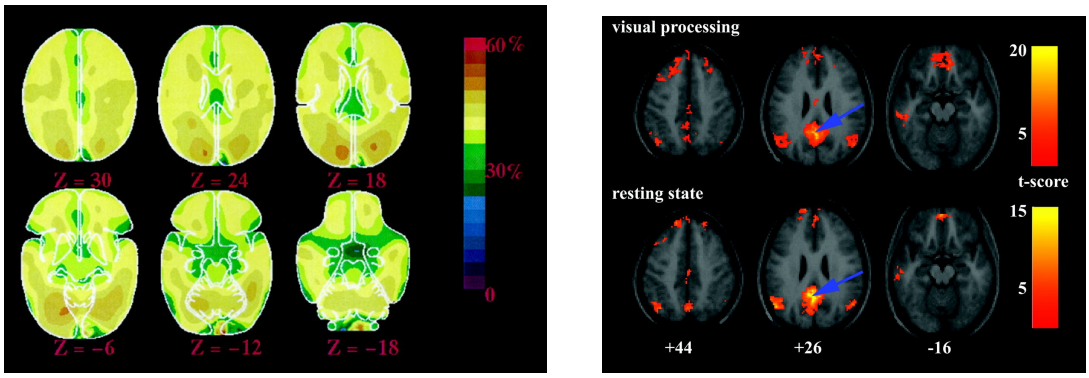

Fig. 1. Left: PET images showing that percentage of available oxygen delivered is relatively constant across the brain during the resting state (subject is lying awake in the scanner, eyes closed). Adapted from Raichle et al. (2001). Right: fMRI images showing that posterior cingulate cortex is functionally connected to the same areas during when subjects view a fixation cross (upper row) and during the resting state with eyes closed (lower row). Adapted from Greicius et al. (2003, Fig. 3).

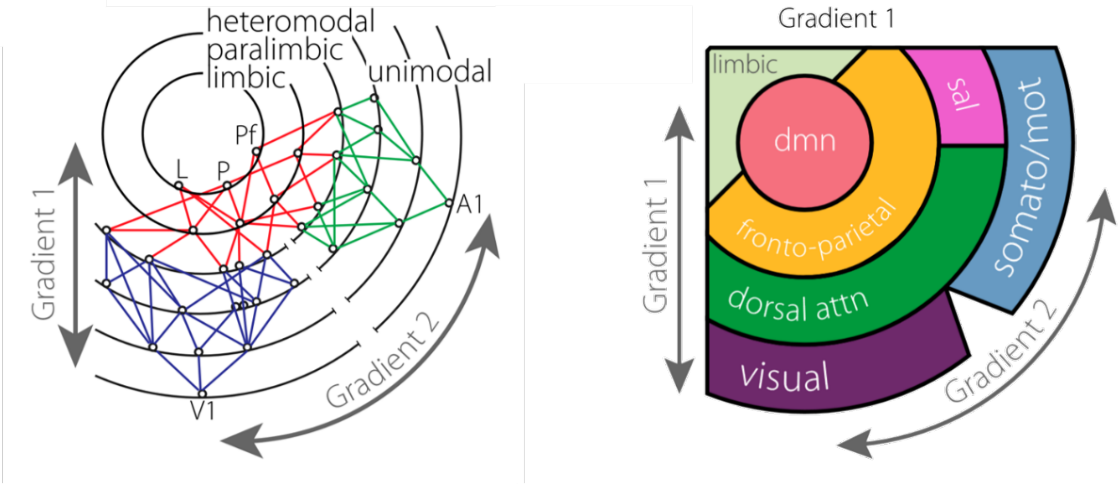

Fig. 2. The Mesulam model (left) is extrapolated from macaque tract-tracing, lesion and electrophysiology data and orders areas according to information represented in unimodal, heteromodal and transmodal areas. The Margulies model is based on diffusion embedding of resting state functional connectivity data (right) situates the DMN at the top of Mesulam's representational hierarchy. Adapted from Margulies et al. (2016).

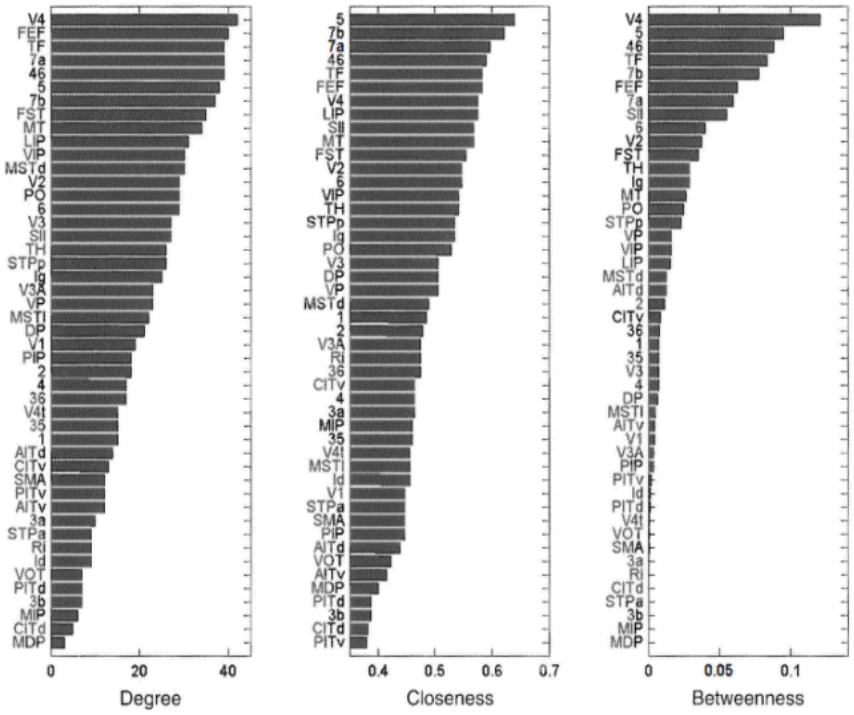

Fig. 3 Measurements of centrality based on Macaque anatomical connectivity data. Degree denotes the percentage of actual out of all possible connections. Closeness centrality is defined as the reciprocal of the sum of the length of the shortest paths between the node and all other nodes in the graph. Betweenness centrality specifies how many shortest paths between any two nodes pass through the node of interest. V4 ranks highest for degree and betweenness centrality and ranks 7th for closeness centrality compared to all other brain areas (adapted from Sporns 2011, Fig. 2.9).

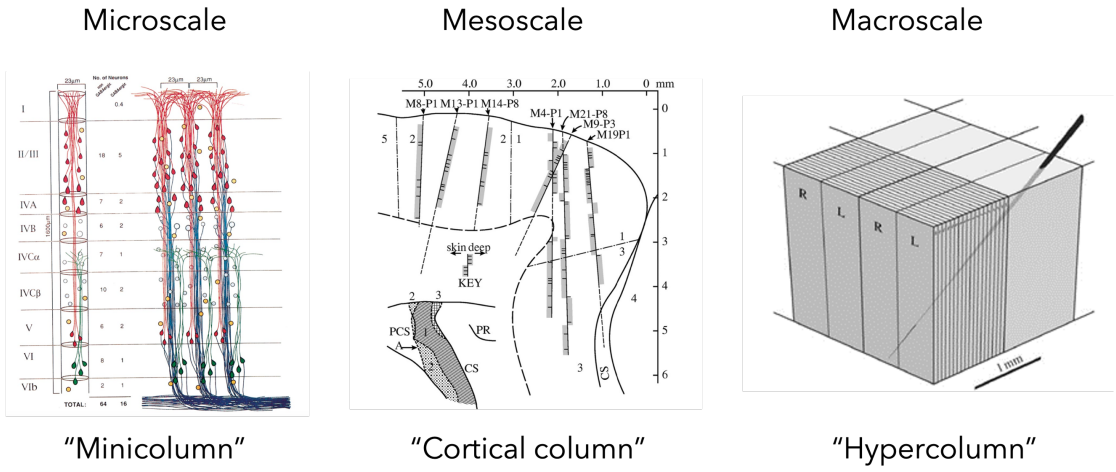

Fig. 4. Different uses of the cortical column concept target different scale-dependent properties of cortical organization. At the microscale, "minicolumn" targets vertical connections of call bands within a 30µm radius (left, adapted from Mountcastle 1997). At the mesoscale, "cortical column" targets functionally modular responses to stimuli in vertical electrode recordings (here: modality-specific responses in cat primary somatosensory cortex that shift every 500µm, adapted from Powell and Mountcastle 1959). At the macroscale, "hypercolumn" targets multiple columnar system which are arranged in a regular fashion (here: sequence-regular orientation columns intersect orthogonally with alternating left and right ocular dominance columns in macaque, this organization is reiterated across entire V1, image from Horton and Adams 2005).

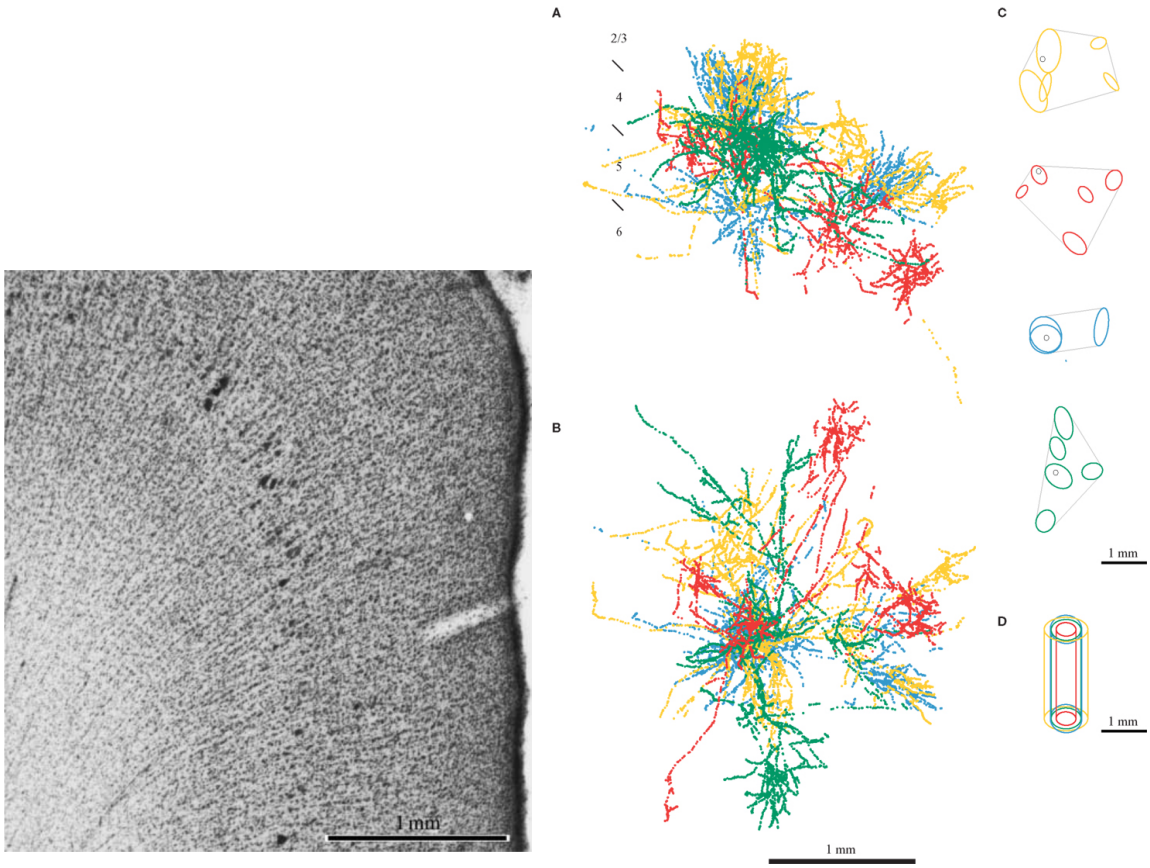

Fig. 5. The problem of missing columnar boundaries. *Left:* Nissl-stained cross-section of macaque somatosensory cortex shows vertical cell bands throughout all cortical layers, but no discernable boundaries between them (adapted from Horton and Adams 2005). *Right:* Stereotypical bouton clustering of neurons in different layers of cat V1 (yellow: L2/3 pyramidal cell, red: L4 spiny stellate cell, blue: L5 pyramidal green: L6 pyramidal cell). In all cases axons extend beyond multiple minicolumns (adapted from da Costa and Martin 2010).

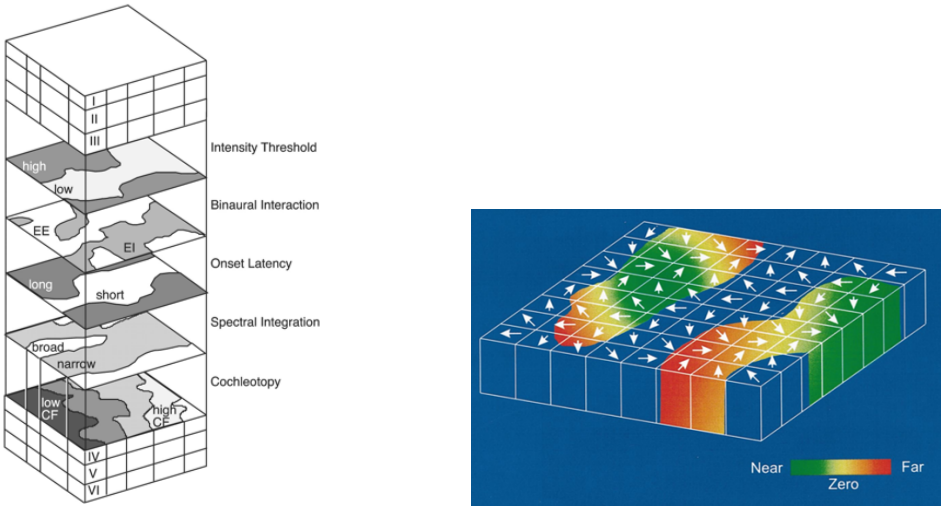

Fig. 6. The problem of non-columnar responses. *Left:* primary auditory cortex contains feature maps, but these do not extend across all cortical layers (adapted from Linden and Schreiner 2003). *Right:* While all MT cells are orientation-selective (white arrows), some are only weakly disparity selective (dark blue regions) Adapted from de Angelis and Newsome (1999).

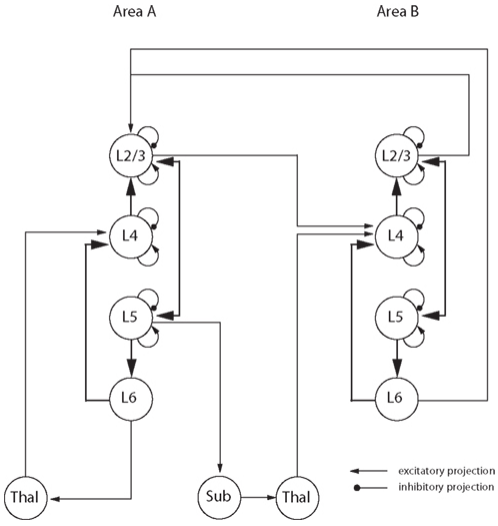

Fig. 7: Wiring diagram of the canonical microcircuit, which is derived from intracellular recordings and horseradish peroxidase staining of neurons in cat V1. Inhibitory and excitatory L4 populations receive weak thalamic input, and send feedforward connections to L2/3 populations that are recurrently connected to L5 populations. L5 populations feedforward to L6 populations and subcortical areas. No spatial dimension is specified, different populations respond differently to stimuli, and the displayed wiring is assumed to be repeated across cortical areas and different species (adapted from da Costa and Martin 2010).

Data availability statement:

No original data were analyzed for this article.

CRediT statement

Philipp Haueis and Daniel S. Margulies equally contributed to the conceptualization of the article. Philipp Haueis led the writing of the original draft, Daniel S. Margulies supported the writing of the original draft.

Funding information

The authors report no funding for this submission.