Cognitive computational neuroscience

J. Brendan Ritchie and Gualtiero Piccinini

Draft - please do not quote or distribute without permission.

The final version is forthcoming in a Bloosmbury volume edited by Nora Heinzelmann

Abstract: *Computational neuroscience consists of building computational models of neural systems at various levels of organization. Most computational neuroscientists assume that nervous systems compute and process information. We explain how computational modelling in neuroscience works by using a recent model of object recognition as a case study and discuss what computational neuroscientists mean by 'computation' and 'information processing' in nervous systems; whether computation and information processing are matters of objective fact or of conventional, observer-dependent description; and how computational descriptions and explanations are related to other levels of analysis and organization.*

1. Introduction

Cognitive computational neuroscience (CCN), also known as computational cognitive neuroscience, is the intersection of cognitive neuroscience and computational (and theoretical) neuroscience. Cognitive neuroscience, in turn, is the study of how nervous systems give rise to cognitive phenomena (Gazzaniga et al. 2019). Computational (and theoretical) neuroscience is the study of phenomena involving nervous systems, including but not limited to cognitive phenomena, by means of mathematical and computational models (Piccinini and Shagrir, 2014; Sejnowski et al. 1988). Thus, CCN is the branch of computational (and theoretical) neuroscience that deals most directly and explicitly with cognition.[1]

[1] A note on terminology may be helpful. Another rough synonym is "model-based cognitive neuroscience," sometimes defined as the "intersection between mathematical psychology and cognitive neuroscience" (Palmeri et al. 2017). "Computational cognitive science" is also used, typically for an enterprise which is less constrained by neuroscience evidence than CCN and is roughly coextensive with "computational psychology". Thus, CCN may also be seen as combining computational cognitive science (or computational or mathematical psychology) with neuroscience (Kriegeskorte and Douglas, 2018; Naselaris et al. 2018). Sometimes, the term "computational cognitive neuroscience" is also used not as a synonym of "CCN" but, more narrowly, for the practice of analyzing neuroimaging data (e.g., fMRI and M/EEG) by means of machine learning methods.

In recent years CCN has branched out from computational neuroscience by establishing its own conference (ccneuro.org, started in 2017) with the aim of identifying the computational processes and principles behind cognition (Naselaris et al., 2018). Computation and representation have long been a touchstone of explanations of mental capacities – at least since McCulloch and Pitts (1943) argued that computations over representations, performed by neural networks, explain cognition (cf. Piccinini, 2004; Colombo and Piccinini, forthcoming). McCulloch and Pitts's computational theory of cognition gave rise to several traditions, including classical cognitive science (e.g., Fodor 1975, Newell and Simon 1972; Pylyshyn, 1984) and connectionism (e.g., Rosemblatt 1958, Rumelhart, McClelland, and PDP Research Group, 1986). It also influenced computational neuroscience. Thus, one may reasonably ask, what distinguishes contemporary CCN from its predecessors? We argue that CCN uses a diverse toolkit of cutting-edge computational models in the service of integrated, multi-level explanation of the neurocomputational mechanisms that underlie cognitive capacities and produce intelligent behavior.

The chapter is structured as follows. In Sect. 2, we sketch the structure of CCN and its ambition of integrated multi-level explanation. While the integrative ambition of CCN is often characterized using Marr's (1982) three levels of analysis for information-processing systems, we argue that the best interpretation of this central ambition is in terms of multilevel neurocomputational mechanisms. In Section 3, we consider a case study on the neural basis of visual object recognition (Hong et al. 2016), which illustrates the application of the CCN toolkit and the sort of mechanistic integration to which it aspires. In Section 4, we address some possible objections to our integrative mechanistic interpretation of CCN. Section 5 concludes the paper.

2. The integrative field of cognitive computational neuroscience

In this section we argue that the ambitions of CCN are well described as the search for neurocomputational mechanisms. First, we contrast CCN with other similar endeavors in the mind sciences and emphasize the importance of its integrated modeling toolkit. While the relationship between CCN modeling approaches is often framed in terms of Marr's (1982) three levels of analysis, characterizing CCN mechanistically makes up for two shortcomings in Marr's framework by providing an account of (i) levels of organization and (ii) computational

implementation. This allows us to characterize the integrative goal of CCN as developing models of multilevel neurocomputational mechanisms that underlie cognition.

2.1 What is distinctive about CCN?

CCN has been described as the intersection of cognitive science, AI, and neuroscience (Figure 1A). But Cognitive Science itself was conceived as an interdisciplinary endeavor that includes connections between psychology, computer science, and neuroscience (Figure 1B). So, in broad interdisciplinary form, the structure of CCN does not so much overlap with, but rather recapitulates, that of Cognitive Science. The stated aim of CCN recapitulates that of Cognitive Science as well. The following passage is a fair characterization of CCN:

> What the subdisciplines . . . share, indeed, what has brought the field into existence, is a common research objective: to discover the representational and computational capacities of the mind and their structural and functional representation in the brain.

However, it comes from the 1978 report to the Sloan Foundation that is often considered one of the documents that first defined Cognitive Science as a field (Keyser et al., 1978, p.6). More generally, the aim of mapping psychological capacities to their neural substrate has historically been the stated objective of *most* attempts at developing a science of the mind.[2] Thus, in both form and ambition, CCN is a reboot of Cognitive Science, or "Cognitive Science 2.0". Has anything changed? We think the answer is yes: methodological innovation in the toolkit of available computational models yields a conception of how an integrated science of cognition might be achieved.

[2] Arguably it is a realization of what Fechner (1860) called "inner" psychophysics. For a comparison between some CCN-style approaches and inner psychophysics, see Ritchie and Carlson (2016). Exceptions include autonomous versions of behaviorism, ecological psychology, and classicist cognitive science. We discuss the autonomy of psychology in Sect. 4.4.

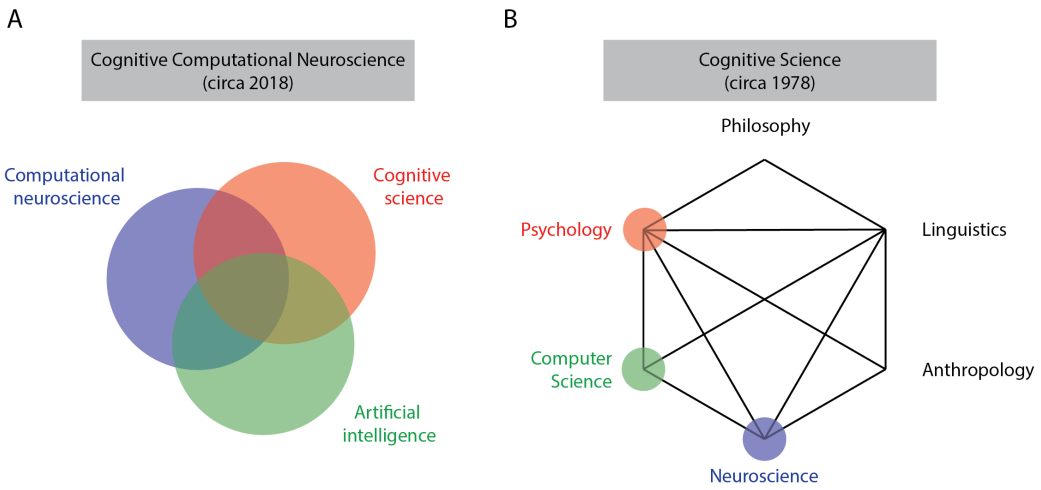

Figure 1: the interdisciplinary structure of CCN and cognitive science as fields. (A) A simplified Venn diagram of the overlapping fields of CCN after Kriegeskorte and Douglas (2018, Figure 2). (B) The different fields of cognitive science and their connections, based on Keyser et al. (1978, Figure 1).

Kriegeskorte and Douglas (2018) characterize CCN as aiming at both cognitive and biological "fidelity" (Figure 2). The main challenge for achieving this goal is building bridges between our theories, as reflected in task-performing computational models, and our experiments, manifested in neural and behavioral data. The methods from different branches of CCN contribute to meeting this challenge in different ways. On one hand, various cognitive neuroscience methods help bridge experiment and theory; these include connectivity models of neural dynamics, neural decoding with machine learning methods, or various "representational" models that characterize the geometry of high-dimensional spaces latent in patterns of brain activity. On the other hand, neural network (NN) models have increasingly been used as a point of comparison to both neural and behavioral data since they can perform similarly to humans, while various formal cognitive models, such as production systems, reinforcement learning, or Bayesian inference characterize rules and norms for how observers perform tasks by decomposing performance into functional units that can be connected to more complex process models (like NNs) and the brain. Taken as a whole, then, while Cognitive Science 1.0 may have faltered in its integrative ambitions (cf. Nunez et al. 2019), its spirit is alive and well under a different name.

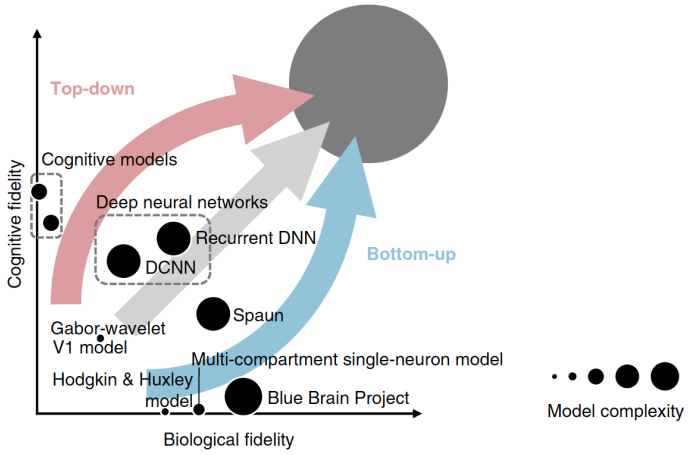

Figure 2. Schematic of the relative cognitive and biological fidelity, and complexity, of different models employed in CCN, from Kriegeskorte and Douglas (2018, Figure 3).

The models highlighted by Kriegeskorte and Douglas (2018) illustrate the impressive toolkit available to CCN, much of which has only become available in the last 20 years.[3] All these different approaches are to be used in concert to address the challenge of relating task-performing models to what brains do. What is the form of this integration? The most common characterization appeals to Marr's three levels of analysis of information-processing systems. At Level 1 is the "computational theory" of a process, which specifies what (mathematical) function is being computed by a system and why it is being computed.[4] Level 2 specifies representational codes of the inputs, internal states, and outputs of a system and an algorithm for achieving the transformation from inputs to outputs. Lastly, Level 3 specifies how the representational and computational capacities of the system are implemented.

Adopting Marr's framework, there might then seem to be a mapping from the three branches of CCN to each level: cognitive science/Level 1, AI/Level 2, and neuroscience/Level 3 (Kriegeskorte and Douglas, 2018; Naselaris et al. 2018). However, this putative correspondence is questionable. In particular, models from AI may map to any of the three levels, depending on what a model is

[3] The formal methods on which the models are based were developed much earlier. It is their large-scale application to the study of the mind and brain that developed more recently, thanks in part to the availability of increased computing power. In this way, methodological and technological developments have helped drive new directions in theory and explanation (Bickle, 2016).

[4] For defenses that the computational theory describes not just the "what" but the "why" of the operation, see Ritchie (2019); Shagrir (2010).

intended to explain (cf. Gentner 2010). This speaks to two more fundamental issues with applying Marr's framework to make sense of CCN and its integrative ambitions. First, the fundamental aim is explanatory, and the models that achieve varying levels of biological or cognitive fidelity are thus related by levels of *organization* not just analysis. As others have pointed out, Marr's framework offers no account of levels of organization or how they may be integrated (Bechtel, 1994; Bechtel and Shagrir, 2015; Zednik, 2017). Second, Marr never specified when the representational schemes and algorithms specified at Level 2 are in fact implemented by a physical system. In other words, he never offered a theory of computational implementation (Ritchie and Piccinini, 2018). Fortunately, both issues can be addressed within a mechanistic framework, and in a way that captures the spirit of CCN.

2.2 CCN and neurocomputational mechanisms

The mechanistic approach aims to capture the form of explanatory practices in many scientific fields – especially biology and neuroscience – which it characterizes as the search for the mechanisms that are responsible for the phenomenon of interest, and thus situate it in the causal structure of the world (Craver and Tabery, 2019). Here we briefly review the core features of mechanistic explanation and emphasize how (i) it affords a multi-level approach to explanation; (ii) computational explanation and implementation can be captured within the framework; and (iii) Marr's levels of analysis are readily explicated within a mechanistic framework.

A mechanism has four elements: (1) some collection of entities – the parts of the mechanism, (2) the causal activities that they perform, (3) the organization of these entities and their activities, and (4) the phenomenon that, under relevant conditions, the mechanism is responsible for (Craver and Tabery, 2019; Illari and Williamson, 2012; Machamer et al., 2000). While "responsible" is a general term, it is also common to speak of a mechanism as producing, underlying, or maintaining a phenomenon (Craver and Darden, 2013). For example, a mechanism produces the phenomenon when thought of as a causal sequence, it maintains a phenomenon when the phenomenon itself is the relative stability of certain variables (e.g., intracellular ion concentrations in cells), and it underlies the phenomenon when it is thought of as a system as a whole. The first two pertain to mechanisms as causal explanations, the third to mechanisms as constitutive explanations.

A mechanistic *explanation* is (a description of) the mechanism that is responsible for the phenomenon.[5] Crucially, mechanisms are multilevel, which is illustrated by how mechanistic explanation often involves multiple iterations of decomposition and localization (Bechtel and Richardson, 2010). If we want to understand how some target system works, we begin by trying to break it down into parts and what they do, how the parts and their activities relate to one another, and how each part and activity contributes to the activities of the whole.[6] A mechanism is at a higher level than entities that make it up, and the further entities that they decompose into are at a lower level, and so on. This notion of mechanistic levels is crucially one of organization, not analysis. Different models may capture different amounts of the complexity of a mechanism, and different levels of organization and levels of abstraction but, ultimately, they are still describing aspects of the same mechanism (Boone and Piccinini, 2016a; Craver and Kaplan, 2020; Glennan, 2005; Milkowski, 2016). This makes sense of how the models from the different branches of CCN might achieve varying levels of cognitive of biological fidelity; they may, to different degrees, capture how a mechanism produces or underlies the phenomenon in question.

The mechanistic approach can also be applied to make sense of the conditions under which a physical system carries out computations, or computational implementation (Kaplan, 2011; Milkowski, 2013; Piccinini, 2007, 2015). According to this account, a physical system that carries out computations is a kind of mechanism that has the function of computing; that is, it's a kind of functional mechanism: the activities of such a mechanism are what their entities are supposed to do (Garson, 2013). For computing mechanisms, the functions of the entities that make up their parts are the computations they perform. Here, the computations are understood to be medium-independent, and whether a physical system implements the computation depends on an appropriate mapping between the different types of vehicles and rules that apply to them that define the computation, and the physical system (Scarantino and Piccinini, 2011; Piccinini, 2015). Whether this is the case depends on whether the physical system has the appropriate degrees of freedom and organization. Crucially, under this account, the only aspects of the physical system that are relevant are those that exhibit the right degrees of freedom and organization, regardless of

[5] Here we set aside whether explanations are best understood as representations of mechanisms (epistemic conception; e.g., Bechtel 2007), as the mechanisms themselves (ontic conception; e.g., Salmon, 1984; Craver 2007), or as a combination of the two (e.g., Boone and Piccinini 2016a).

[6] We return to the topic of localization when considering objections in Section 4.

the nature of the medium. Thus, if CCN aims to provide models that capture the computational principles that underlie our mental capacities, the mechanistic approach provides an account of the conditions under which the brain might implement the operations described by these models.

Not only does the mechanistic approach flesh out the details lacking in Marr's levels of analysis, it also can be characterized in line with Marr's framework. At level 1, the why-component includes functional, non-mechanistic details that constrain what functions might be suitable given the task that is being carried out (i.e., the phenomenon). Thus, here we have larger contextual details related to the function and environmental context of the system (Bechtel, 2009; Bechtel and Shagrir, 2015). At Level 2, we have a specification of how the information is encoded and of the medium-independent computational processes that describe how the formal operations at Level 1 are in fact carried out. This level presumes that the decomposition is of the actual operations that are being performed by the system, which are underdetermined by the details present at Level 1. At Level 3, we have the implementation of the computational processes and representational code at Level 2, which specifies the localization of the different entities and activities in the target system (in this case the brain). The explanation iterates so that, at every level of organization, we can give a computational, algorithmic, or implementation account. The computations performed by lower-level components compose the computations performed by higher-level components, until we provide a mechanistic explanation of the behavior of the whole system. The mechanistic account of computation is intended precisely as an account of how computations are implemented in physical systems. The relevant details are such that the entities that make up the physical system exhibit the right degrees of freedom to carry out the computation or compose descriptions from a representational code.

In summary, to the extent that CCN aims to provide models of the computational principles that underlie our mental capacities, and the models are to be evaluated based on their cognitive and biological fidelity, then this picture is readily captured by the mechanistic approach to explanation (Boone and Piccinini, 2016b). To see how this all looks in practice, we turn to a case study.

3. A case study in CCN: the neural basis of visual object recognition

So far, we have discussed CCN and mechanistic explanation in fairly abstract terms. Here we consider a detailed case study that illustrates why the sort of theoretical integration CCN aims for is well-described in terms of multilevel explanation of neurocomputational mechanisms. Specifically, we focus on the study of Hong et al. (2016) on the representation of category-orthogonal information along stages of the ventral visual pathway in the primate brain. After reviewing the motivation, design, results, and interpretation of the study, we consider the different ways in which it illustrates the kind of mechanistic integration that we believe CCN aspires to.

3.1 Representing category-orthogonal information in the ventral visual stream

Most models of the visual system maintain that the later stages of the ventral visual pathway represent more abstract object properties like category membership ("cat") and the identity ("this particular cat") of individual exemplars (DiCarlo et al. 2012; Riesenhuber and Poggio, 2000). However, in natural viewing conditions we never see the same object under identical conditions; there is always a change in viewpoint related to position in the visual field, orientation, distance, illumination, and so forth. Thus, somehow the visual system must overcome this "invariance problem" by forming representations that are, on one the hand, tolerant to this sort of variation and, on the other, still specific to target identities and categories (Pinto et al. 2008; Rust and Stocker, 2010).

The study of Hong et al. (2016) addresses the flipside of this issue: how does the visual system represent the visual properties that constitute such variation in viewpoint; that is, how do stages of the ventral pathway represent visual properties that are orthogonal to object identity and category membership (Bracci et al. 2017)? They propose four hypotheses based on known properties of the visual system, which they couch in terms of how well category-orthogonal information can be decoded from patterns of neural activity at each stage (Figure 3): (1) the decodability of the category-orthogonal object properties is at or above human performance in early stages and decreases; (2) decodability is greater at intermediate stages; (3) decodability is at human level performance across all stages; and (4) decodability increases along the ventral stream just as it does for classifying object category and identity.

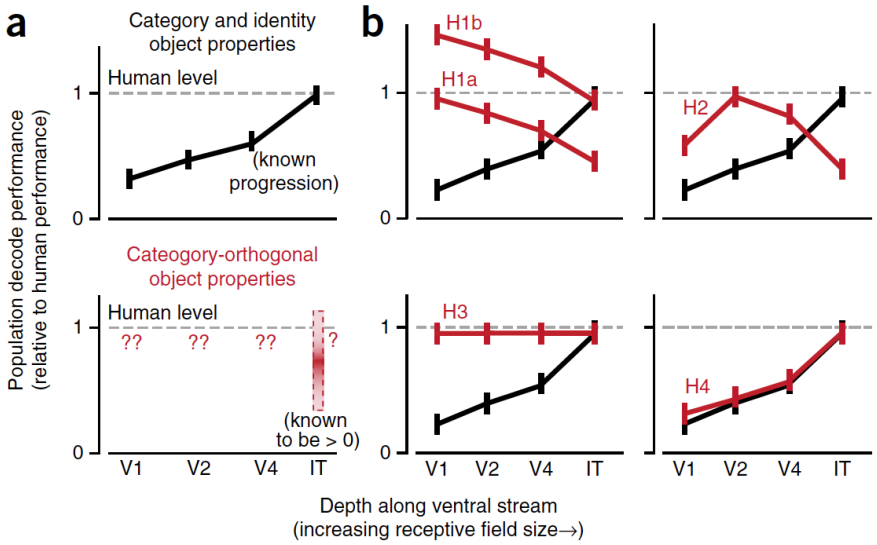

Figure 3. Hypothesis tested by Hong et al. (2016). X-axes indicate stages in the macaque ventral visual pathway. Y-axes indicate (hypothetical) proportion of accurate classification of images by a machine learning classifier trained to guess image labels from neural recordings at each stage as a proportion of human accuracy when categorizing the same images. (a) Known proportion performance when classifiers label images based on object category and identity, and schematic indicating the unknown status of decodability for category-orthogonal object properties. (b) Four hypotheses regarding the decodability of category-orthogonal information at each stage of the ventral pathway.

To evaluate these alternative hypotheses, Hong et al. measured neural population responses from two stages of the ventral pathway in rhesus macaques (N = 2): visual area 4 (V4) and inferior temporal (IT) cortex, which encompasses the final stages of the visual processing stream.[7] In place of V1 or V2 (which were not recorded from), the early layers of a pretrained deep neural network (DNN) were used as a proxy model for early visual cortex. Stimuli consisted of a large number of images containing objects of 8 categories (animals, boats, cars, chairs, faces, fruit, planes, and tables) positioned in front of natural scene images with the specific position varying in many different properties that were orthogonal to category membership, such as position, shape, bounding box, orientation, and rotation. Even with such variation, humans perform extremely well at categorizing the objects in the stimulus set.

[7] Recordings were made from 9 chronically implanted electrodes in 3 hemispheres, with a total of 266 sites in IT and 126 in V4.

In their analysis, the researchers first compared category and visual property representations across the brain areas. They found that at the single site level, the best individual IT sites contain more discriminable information about category, or their activity correlated more with individual orthogonal visual properties, than V4 sites for most of the visual properties. Because information for these modeling tasks is often spatially distributed across multiple neuronal sites, they next tried to decode category and visual property information from all of the sites in V4 and IT (Figure 4). In the case of the categorization task, they used a support vector machine classifier and, for the visual properties, an L2-regularized linear regression model. They found that, for all tasks, IT revealed greater decodable information than V4. When compared to a pixel-wise model and the proxy for V1 from the pretrained DNN, IT had significantly more decodable information than either model, while V4 had more decodable information for some but not all of the tasks. Notably, similar analysis using grating stimuli showed the opposite pattern, with far greater decodable information for visual properties in horizontal and vertical position and orientation in the model of V1. This was an important control condition, since gratings properties are known to be decodable from early visual cortex and can also vary in some of the same category-orthogonal properties manipulated in the main experiment.

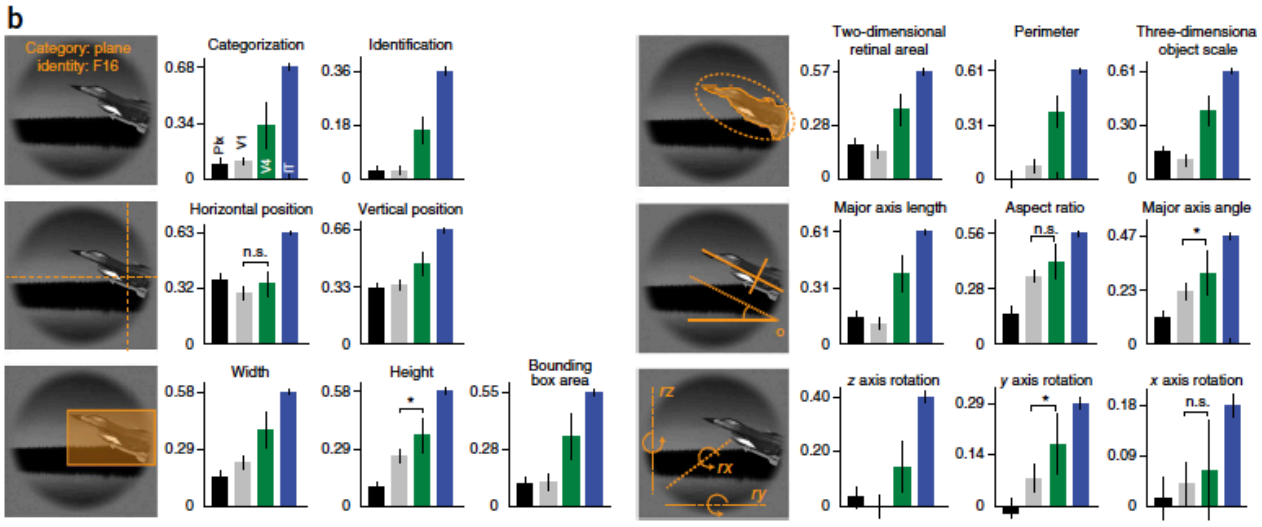

Figure 4. Population decoding results from Hong et al. 2016. Y-axis is linear classifier accuracy for the different tasks.

Next, the authors collected behavioral data on a subset of the tasks, including categorization, position, size, pose, and bounding-box estimation. They sought to determine the number of neural sites needed to achieve performance level with decoding equivalent to humans. For each brain area they took different sized samples of sites and ran linear decoders as before. They consistently found that, as number of sites increased, human-level performance would be achieved with a fewer number of sites when sampling from sites in IT than V4, and for sites from V4 than for the V1 model and pixel models.

The researchers investigated the distribution of decodable information across IT sites and asked how much overlap there might be in the recording sites in terms of the source of decodable information for the different tasks or whether these tasks recruit different subpopulations of neurons. To do this they looked at the classifier weights as a proxy for how important a site is to a task, with positive weights being treated as important for the task. Then, for each distribution of task weights, they measured sparseness and imbalance. Sparseness indicates the proportion of total sites informative for a task, while imbalance is the proportion of sites correlated (rather than anticorrelated) with the task. They found that across tasks ~26 % of sites were highly weighted and about half had indistinguishable sparseness from a normal distribution, while, for most tasks, imbalance was not different from chance. To quantify overlap, they calculated the correlation of the absolute values of the decoder weight vectors for each task pair. The proportion of overlap determined the degree to which downstream stages could use the same set of sites to perform the pair of decoding tasks. Across all pairs > 50 % had significantly positive overlap, and 16.6% had negative overlap. While related tasks tended to have more overlap, even unrelated tasks had more overlap than one would expect from chance. These results suggest that IT neurons jointly code for both categorical and category-orthogonal information.[8]

As DNNs have been used as proxy models for the primate ventral pathway when it comes to object categorization, the authors also assessed how a model trained to classify object categories (separate from the ones used in the study) would represent the target object categories and orthogonal visual properties. Consistently with prior work, they found that the top hidden layers were more predictive of the neural response in IT and intermediary layers were most predictive of responses

[8] Faces were the one possible exception to this trend (Hong et al. 2016, p.617-618).

in V4, while the early layers also showed a V1-like Gabor edge tuning. Task performance was also evaluated for each layer. They found that, over the time course of training, performance in the top hidden layer improved for category-orthogonal information. This despite the fact that the model was being trained on categorization, and not the visual estimation tasks. Thus, the trained DNN exhibited similar representational architecture across layers as was observed for the stages of ventral pathway processing.

Taken as a whole, the pattern of results reported by Hong et al. 2016 provides clear support for hypothesis (4): as with category membership, category-orthogonal properties are increasingly decodable along the processing stream of the ventral pathway in the monkey brain. Thus, in the course of building up viewpoint tolerant and category-specific representations, the visual system also builds up more specific representations of category-orthogonal properties from the visual input.

3.2 Computational mechanistic integration and the ventral visual stream

There are a number of ways in which the study of Hong et al. nicely illustrates CNN in practice and weaves together theories and models from cognitive science, AI, and neuroscience with different degrees of cognitive and biological fidelity. This can be seen by briefly considering some of the subtext to the project. On one hand, the framing in terms of viewpoint invariance reflects a foundational characterization of object recognition from visual psychophysics (Biederman, 1987; Marr and Nishihara, 1978). On the other hand, the application of DNN models and machine learning methods to the distribution of recording sites is driven by the idea that the brain implements population codes in which the vehicles for representational content are distributed patterns of neural activity (Ebitz and Hayden, 2021; Jazayeri and Afraz, 2017; Pouget et al. 2000), which is also a feature of DNN architectures in AI that are characterized by distributed representations (Hinton et al. 1986; LeCun et al. 2015).

The Hong et al. study is also a clear example of the mechanistic approach at work, since it is an investigation of the mechanisms that underlie and produce object recognition. We see this because it follows a strategy of decomposition and localization for the relevant components and their activities in the ventral pathway. The pathway itself consists of four major stages (V1, V2, V4,

and IT) and the researchers worked off a sketch of how these stages are organized: a (primarily) feedforward pathway with each stage engaging in the activity of representing different information extracted from the visual input and IT representing information about object category and identity (and, as their study shows, category-orthogonal properties as well). The study then considers four different "how-possibly" models as to what these representational capacities are like, in terms of which stage contains the most decodable information. The analyses that are carried out are then intended to test the decodable information for category and category-orthogonal tasks in the neural measures. As a whole the project illustrates how the mechanistic approach captures the multi-level and computational aspects of CCN.

When it comes to levels of organization, the models from the different fields are related based on degrees of abstraction and detail. The high-level models of object recognition sketch mechanisms that can be mapped onto more detailed models with specific neural components of the ventral pathway via theories of the computational encoding schemes. Thus, the relative biological and neural fidelity of the different models relied on by Hong et al. all describe the mechanisms with different degrees of abstraction, which characterize different levels of the same mechanism. For example, the different how-possibly models specifically relate to the activities at one mechanistic level, but the sparsity of the coding for the information related to the different tasks happens at a lower mechanistic level.

Regarding computation, there is a particular conception of neural information processing that is presumed by the study, that of a neural population code (Jazayeri and Afraz, 2017; Pouget et al. 2000). Under this picture, the vehicles for content are distributed patterns of neural activity and the rules for extracting this information roughly conform to that of a linear read-out by later processing stages (Laakso and Cottrell, 2000; Shea, 2007). So, the brain is implementing a particular kind of information-processing architecture because we can map a distributed representational code to the brain and the brain has the function of carrying out the operations over this code (Williams and Colling, 2018). It is also straightforward to describe the project as having elements at all three levels of Marr's hierarchy (Grill-Spector and Weiner, 2014; Ritchie, 2019), since it relies on a specification of what the visual system is trying to do (viewpoint invariant object

recognition), posits a representational format (distributed coding), and its implementation (population code).

We have chosen to focus on the study of Hong et al. because it is a particularly compelling example of both CCN and how it can be captured in the mechanistic framework. It is also representative of a large amount of recent and ongoing research that brings together the different branches of CCN to investigate the neurocomputational mechanisms that underlie our mental capacities.

4. Objections to the mechanistic characterization of CCN

So far, we have made the case that CCN aims at multi-level explanations of the neurocomputational mechanisms that underlie psychological capacities and produce behavior. In this section we consider a number of reasons one might doubt that CCN should be characterized in this way, based on various critiques of the mechanistic approach to explanation. We also consider whether CCN is fundamentally misguided, since it aspires to a form of explanatory integration between computational models that cannot be achieved if psychology is autonomous from neuroscience.

4.1 The CCN modeling toolkit is diverse, but the mechanistic approach is hegemonic

CCN is supposed to have an impressive modeling toolkit drawn from cognitive science, AI, and neuroscience. A crucial feature of this toolkit is its *diversity*: different modeling approaches describe different aspects of our psychological capacities or the inner workings of the brain at some scale of function and organization. However, many have claimed that the mechanistic approach tends to assimilate different modeling approaches into a single, narrow perspective that downplays, or discounts, diversity in the explanatory goals of different modeling approaches. Thus, while the toolkit of CCN is diverse, the mechanistic approach is *hegemonic* (Shapiro, 2017).

Many different modeling approaches have been contrasted with modeling mechanisms, including functional analysis (Shapiro, 2017; Weiskopf, 2011), dynamical systems modeling (Barack, 2021; Meyer, 2020; Ross, 2015; Silberstein and Chemero, 2013), network analyses (Huneman, 2010; Levy and Bechtel, 2013; Rathkopf, 2018), pathway models (Ross, 2018, 2021), Bayesian modeling (Rescorla, 2018), and computational and mathematical modeling more generally (Batterman and

Rice, 2014; Chirimuuta, 2014). Mechanists have replied that applications of these modeling approaches get their explanatory *force* to the extent that they map onto the elements of underlying mechanisms (Craver, 2016; Kaplan, 2015; Kaplan and Craver, 2011; Piccinini and Craver, 2011; Povich, 2015; Zednik, 2018). Recall that the mechanistic approach describes how explanations situate phenomena of interest in the causal structure of the world; that is, it is an account of *constitutive* causal explanations. To the extent that these or other modeling approaches are used for other purposes, they need not describe mechanisms as such (Chirimuuta, 2018; Hochstein, 2017; Craver and Kaplan, 2020). In addition, mechanistic explanation must also draw on broad contextual factors when explain cognitive capacities, as illustrated by our earlier discussion of Marr (cf. Bechtel, 2009; Shagrir and Bechtel, 2015, Zednik 2017, Fuentes Muñoz forthcoming).[9]

One rejoinder is that the above response ignores a deeper incongruency, which is that even when we restrict ourselves to constitutive causal explanation, the mechanistic approach is overly restrictive since it requires that we follow norms of decomposition and localization (Bechtel and Richardson, 2010). However, the crucial feature of many of the above approaches (e.g., functional analysis, dynamical systems modeling, and network analysis) is that they violate these norms when it comes to how they map onto the brain.

This rejoinder presupposes a strong reading of the strategy of decomposition and localization. But decomposition and localization can be understood more weakly (Burnston, 2021; cf. Dewhurst and Isaac, 2021). On a strong reading, each cognitive function admits of a unique decomposition into subfunctions, each subfunction is carried out by a unique neural structure, and each structure carries out one and only one subfunction. A reductionist assumption sometimes included in the strong reading of localization is that individual subfunctions can be studied and fully understood in isolation from one another (Kaiser and Krickel, 2017). For better or worse, there is plenty of evidence that many neurocognitive functions do not admit of decomposition and localization in this strong sense (Anderson, 2014; Burnston 2021, Pessoa 2022, McCaffrey forthcoming). On a weak reading, cognitive functions may be decomposable in more than one way depending on

[9] In this respect, appealing to causal interventions as a way to try and show that dynamical models (Meyers, 2020) or cognitive models (Rescorla, 2018) are not mechanistic is ill-considered. Intervention is always carried out with respect to the causal structure of the world, which is what mechanistic explanation is after.

context, each subfunction may be carried out by different neural structures depending on context, neural structures may carry out different subfunctions depending on context, and subfunctions may be fully understandable only when studied in combination with other subfunctions. Modeling approaches that are contrasted with mechanistic explanation violate only the strong construal of decomposition and localization. But a mechanistic approach to CCN need only be committed to the weak reading of this strategy. When a weak reading is accepted, insofar as other modeling approaches contribute to constitutive causal explanation of cognitive capacities, they describe aspects of neurocognitive mechanisms.

So, when characterized in a manner that is pluralistic about the explanatory roles of modeling approaches and combined with a weak construal of decomposition and localization, our mechanistic characterization is readily compatible with the diversity of modeling approaches in CCN.

4.2 Mechanistic explanation demands biological fidelity at the expense of cognitive fidelity

The mechanistic approach is supposed to account for how CCN aims for models that exhibit varying degrees of both cognitive and biological fidelity, which it characterizes in terms of describing the levels of mechanistic organization that produce and underlie our cognitive capacities. However, achieving cognitive fidelity often requires models that provide "high-level" descriptions of mental processes that abstract away from implementation details. In contrast, many have argued that the mechanistic approach is inconsistent with this sort of abstraction, since, generally, the more detailed a description of a mechanism, the better the explanation. Indeed, this conflict with abstraction and demand for details is often a primary reason why the sorts of modeling approaches canvased earlier are not mechanistic (Batterman and Rice, 2014; Chirimuuta, 2014; Ross, 2021; Weiskopf, 2011). Thus, it would seem that the mechanistic approach asks that we sacrifice cognitive for biological fidelity.

There are two aspects to this objection. The first is that mechanistic explanation is somehow in conflict with abstraction in modeling. This is simply not the case (Boone and Piccinini, 2016a; cf. Levy and Bechtel, 2013). There are many forms of abstraction consistent with models providing partial (principled) descriptions of mechanisms, including but not limited to abstractions due to

selection of a relevant level of organization, idealization, incomplete knowledge, or mathematical convenience. Even if this is conceded, however, there remains the separate question whether the mechanistic approach claims that "more details are better" when it comes to modeling. Whatever the nature of earlier formulations, this is not a requirement of the mechanistic approach, for the same sorts of reasons alluded to in response to the above objection (Craver and Kaplan, 2020; Milkowski, 2016). Briefly, no mechanism can be described in full detail by any model, which inherently must involve idealization and abstraction. So, it is simply not the case that greater detail always improves the explanatory import of a model. Additionally, mechanistic modeling, like any kind of modeling, requires perspective taking (Kastner, 2018; Lee and Dewhurst, 2021). Thus, even mechanistic models must be developed at the level of organization that is relevant to explaining a phenomenon of interest, which requires selecting the variables that are most relevant at that level and omitting lower-level details that make little or no difference to the phenomenon. Instead, all else being equal, *amalgams* of models collectively give us more complete explanations of the phenomena we are interested in, although each will only describe some fragment of the mechanism that produces or underlies the phenomenon (Craver and Kaplan, 2020).

The foregoing holds especially when it comes to models of the mechanism that *underlies* a phenomenon of interest. However, it may be that the ambitions of CCN do entail a commitment to one kind of complete model, namely one that exhibits behavior that is equivalent to that of humans performing some cognitive task (Kriegeskorte and Douglas, 2018). This does conform to one sense of completeness for mechanistic models: that they produce the phenomenon, within a certain range of conditions (Baetu, 2015). It is important to emphasize that even in this case mechanistic models still involve a great deal of idealization and simplification. For example, the SPAUN model of Eliasmith et al. (2012) is intended to model a complete brain simplified to 2.5 million neurons allocated to different cognitive processes based on their neural implementation (e.g., information encoding of visual input, reward evaluation, action selection, working memory, and motor processing). This model aims for completeness in the sense of exhibiting behavior across a number of cognitive tasks while exhibiting both cognitive and biological fidelity at a coarse scale of description.

4.3 Computation is not mechanistic

The other crucial aspect of our mechanistic characterization of CCN is that it accounts for how a physical system can implement a computation. However, the mechanistic account of computation is controversial and certain critiques are particularly germane to the present discussion.

One concern is how mechanistic levels of organization are even supposed to relate to the implementation relation (Coelho Mollo 2018; Elber-Dorozko and Shagrir, 2021). On one hand, mechanisms at different levels stand in part-whole relations with respect to each other and the phenomenon that they explain. On the other hand, the implementation relation applies between an abstract, medium-independent description and a physical system while exhibiting its own compositional organization often characterized in functional terms. The question, then, is how the mechanistic and computational hierarchies can be related so that computational explanation is mechanistic.

We believe this concern is more apparent than real, for two reasons. First, as argued by Piccinini (2020), the purely computational (i.e., medium-independent) and the implementational (i.e., medium-dependent) hierarchies of organization are both mechanistic. Which of the two we choose to articulate depends on what we are trying to explain. This coheres with the holistic perspective on modeling referenced earlier. Different computational models may relate to either medium-independent, or dependent, decompositions of the same system. Earlier we noted that the study of Hong et al. made use of DNN models that characterize representations of object properties as a code distributed across nodes in network layers while at the same time applying machine learning classifiers to pattern of responses at recording sites based on the hypothesis that the brain codes for these properties at the population level. Both of these models of the representational mechanisms for object recognition, but one is purely computational and the other implementational.

Second, the distinction between computation and implementation mirrors the more general distinction between functional roles and their realizers (Ritchie and Piccinini, 2018). In this respect, it is important to not fall prey into the sort of "two-levelism" thinking that is invited by classic distinctions between structure/function and hardware/software that are so intuitive when discussing physical computation. As observed by Lycan (1987), the distinction between role and

realizer is *relative* to a particular level in an organizational hierarchy. For example, a population of neurons will implement a distributed representation of object properties in virtue of the computations they individually perform, so the implementation for the code is itself the result of the computational operations of the cells. So, at a particular computational level, the corresponding implementation is constituted by lower levels of computation. When this sort of relation between computational and implementational levels is kept in mind, then there is again no mystery as to how the computational and implementational hierarchy are related.

Another, more basic concern, is that the mechanistic account of computational implementation is inadequate. Recall that under this view, a physical system implements a computation when it is a mechanism that meets mapping requirements to the organization of the computational description and has the function of carrying out the computation in question. We do not have the space here to canvas the adequacy of this mechanistic approach.[10] Whether or not this succeeds as a *general* account of implementation, in the present context it suffices that it holds in a more restricted form. If our cognitive capacities are explained mechanistically and computationally, then, per our response to the previous concern, the medium-independent, computational structure will satisfy conditions for the description of a mechanism. Put simply, if we have independent reason to think the system in question is a mechanism that has the function of carrying out certain computations, then the mechanistic account applies.[11]

4.4 Psychology is autonomous

A different sort of objection targets the explanatory ambitions of CCN directly. Many aspects of the traditional branches of cognitive science have arguably become disunified (Nunez et al. 2019). Researchers may see no obligation to clamor on this scientific bandwagon. An even stronger claim is that the fundamental ambition of CCN is misguided. Built into the idea that we can develop integrated, multilevel explanations of cognitive phenomena is the idea that the different branches of CCN are engaged in the same explanatory enterprise. However, some philosophers of

[10] For some recent discussion, see Chirimuuta (2022), Kirkpatrick (2022), Maley (2022), Piccinini (2022), Shagrir (2022).

[11] This response might not be available if one adopts a non-functional version of the mechanistic view of implementation, according to which what operations a system carries out is determined solely by the intrinsic properties of a system (Dewhurst, 2018; Coelho Mollo, 2018).

psychology have resisted this perspective, instead seeing it as largely "autonomous" from neuroscience in particular.

There are many senses in which psychology may be autonomous (Piccinini, 2020). Some of these senses are readily accommodated by a mechanistic perspective. For example, if autonomy is the claim that psychology does not reduce to neuroscience (Fodor, 1975, 1997), then our mechanistic characterization of CCN supports autonomy.[12] Similarly, if autonomy is simply the claim that psychology and neuroscience offer us different kinds of mutually constraining models, as part of a larger explanatory enterprise, then it is likewise compatible with CCN (cf. Hochstein, 2016). But a sense that is *not* compatible with CCN is that psychology and neuroscience are engage in distinct explanatory enterprises that may proceed independently of one another; in particular, that cognitive models can be confirmed or disconfirmed without reference to the sorts of facts typically revealed by neuroscience (Barrett, 2014; Shapiro, 2017; Weiskopf, 2011, 2017).

This sort of view flies in the face of how behavioral evidence is used in practice to critically evaluate models from neuroscience based on their cognitive fidelity, and in turn, to evaluate the explanatory import of the models themselves (Povich, 2015). To see this, first consider a different thesis: that cognitive models in psychology are autonomous from AI. Such a position is hard to reconcile with how facts from psychophysics and cognitive models are used to evaluate the capacities of different architectures, such as the performance of deep neural networks on visual tasks (Bowers et al. 2022; Serre, 2019). These critiques are only intelligible if we are evaluating DNNs as models of the *very same* phenomena that cognitive models from psychology hopes to explain. The same considerations apply when it comes to how behavioral measures and cognitive models are used *within* neuroscience. For example, many have pointed out that behavioral measures are necessary for neuroscience precisely because of how they are revealing of the cognitive capacities of the neural systems being investigated (Krakauer et al. 2017; Niv, 2021). Similarly, so-called "model-based" cognitive neuroscience focuses on relating parameters of cognitive models of behavior to neural signals in order to illuminate the behavioral relevance of

[12] For this reason, bringing up familiar points about multiple realization, which are intended to rebut reductionism, as an argument against explanatory integration, misses the point. See Piccinini (2020) for discussion.

neural function (Turner et al. 2018; Povich, 2015; Ritchie and Carlson, 2016). In short, cognitive models are part of the CCN toolkit precisely because they are indispensable to neuroscience.

In the other direction, this sort of view flies in the face of psychology's ambition to discover not just a way that cognition might work, but the actual way that cognition takes place. Insofar as the computational models posited by psychologists and classical cognitive science are accurate, the computations they posit are carried out by actual neurocomputational mechanisms. To show that they are, cognitive models must be embedded within neurocomputational models posited by CCN. Another way to put this point is that the neurocomputational machinery uncovered by CCN must be able to generate the computations posited by psychological models on pain of the latter being relegated, at best, to mere how-possibly explanations (Piccinini 2020; Ritchie and Piccinini, 2018).

5. Conclusion

At root, CCN is the branch of computational (and theoretical) neuroscience that most directly relates to cognition. We have argued that what is distinctive of CCN is how it draws on a rich modeling toolkit with the aim of developing models of the neurocomputational mechanisms that underlie our cognitive capacities and produce intelligent behavior. We illustrated this mechanistic characterization of CCN using a case study on how the visual system represents category-orthogonal object properties (Hong et al., 2016). Finally, we defended our characterization from a number of objections often directed at mechanistic approaches to explanation. Addressing these objections helped to further reveal that the mechanistic ambitions of CCN open the door to diverse modeling practices when explaining the neural basis of cognition.

References

Anderson, M. L. (2021). *After phrenology: Neural reuse and the interactive brain*. MIT Press.

Baetu, T. M. (2015). The completeness of mechanistic explanations. *Philosophy of Science*, *82*(5), 775-786.

Barack, D. L. (2021). Mental kinematics: dynamics and mechanics of neurocognitive systems. *Synthese*, *199*(1), 1091-1123.

Barrett, D. (2014). Functional analysis and mechanistic explanation. *Synthese*, *191*(12), 2695-2714.

Batterman, R. W., & Rice, C. C. (2014). Minimal model explanations. *Philosophy of Science*, *81*(3), 349-376.

Bechtel, W. (1994). Levels of description and explanation in cognitive science. *Minds and Machines*, *4*(1), 1-25.

Bechtel, W. (2007). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Psychology Press.

Bechtel, W. (2009). Looking down, around, and up: Mechanistic explanation in psychology. *Philosophical Psychology*, *22*(5), 543-564.

Bechtel, W., & Richardson, R. C. (2010). *Discovering complexity: Decomposition and localization as strategies in scientific research*. MIT press.

Bechtel, W., & Shagrir, O. (2015). The non-redundant contributions of Marr's three levels of analysis for explaining information-processing mechanisms. *Topics in Cognitive Science*, *7*(2), 312-322.

Bickle, J. (2016). Revolutions in neuroscience: Tool development. *Frontiers in systems neuroscience*, *10*, 24.

Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, *94*(2), 115.

Boone, W., & Piccinini, G. (2016a). Mechanistic abstraction. *Philosophy of Science*, *83*(5), 686-697.

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

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., ... & Blything, R. (2022). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, 1-74.

Bracci, S., Ritchie, J. B., & de Beeck, H. O. (2017). On the partnership between neural representations of object categories and visual features in the ventral visual pathway. *Neuropsychologia*, *105*, 153-164.

Burnston, D. C. (2021). Getting over atomism: Functional decomposition in complex neural systems. *The British journal for the philosophy of science*.

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

Chirimuuta, M. (2018). Explanation in computational neuroscience: Causal and non-causal. *The British Journal for the Philosophy of Science*.

Chirimuuta, M. (2022). The Case for Medium Dependence. *Journal of Consciousness Studies* 29 (7-8): 185-194.

Colombo, M., and Piccinini, G. (forthcoming). *The Computational Theory of Mind*. Cambridge: Cambridge University Press.

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

Craver, C. F. (2016). The explanatory power of network models. *Philosophy of Science*, *83*(5), 698-709.

Craver, C. F., & Darden, L. (2013). *In search of mechanisms: Discoveries across the life sciences*. University of Chicago Press.

Craver, C. F., & Kaplan, D. M. (2020). Are more details better? On the norms of completeness for mechanistic explanations. *The British Journal for the Philosophy of Science*.

Craver, C., & Tabery, J. (2015). Mechanisms in science. *Stanford Encyclopedia of Philosophy*.

Dewhurst, J. (2018). Computing mechanisms without proper functions. *Minds and Machines*, *28*(3), 569-588.

Dewhurst, J., & Isaac, A. (2021). The ups and downs of mechanism realism: Functions, levels, and crosscutting hierarchies. *Erkenntnis*, 1-23.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415-434.

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

Eliasmith, C., Stewart, T. C., Choo, X., Bekolay, T., DeWolf, T., Tang, Y., & Rasmussen, D. (2012). A large-scale model of the functioning brain. *science*, *338*(6111), 1202-1205.

Fechner, G. T. (1860). *Elemente der psychophysik* (Vol. 2). Breitkopf u. Härtel.

Fodor, J.A. (1974) Special sciences (or: The disunity of science as a working hypothesis). *Synthese.* 28, 97–115.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Fodor, J. (1997). Special sciences: Still autonomous after all these years. *Philosophical perspectives*, *11*, 149-163.

Fuentes Muñoz, J. I. (forthcoming). Efficient Mechanisms. *Philosophical Psychology*.

Garson, J. (2013). The functional sense of mechanism. *Philosophy of science*, *80*(3), 317-333.

Gentner, D. (2010). Psychology in cognitive science: 1978–2038. *Topics in Cognitive Science*, *2*(3), 328-344.

Glennan, S. (2005). Modeling mechanisms. *Studies in history and philosophy of science part C: studies in history and philosophy of biological and biomedical sciences*, *36*(2), 443-464.

Grill-Spector, K., & Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, *15*(8), 536-548.

Hinton, G.E., McClelland, J.L., and Rumelhart, D.E. (1986). Distributed Representations. Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing*: *Vol. 1*. Cambridge, MA: MIT Press/Bradford Books, 77-109.

Hochstein, E. (2016). Giving up on convergence and autonomy: Why the theories of psychology and neuroscience are codependent as well as irreconcilable. *Studies in History and Philosophy of Science Part A*, *56*, 135-144.

Hochstein, E. (2017). Why one model is never enough: a defense of explanatory holism. *Biology & Philosophy*, *32*(6), 1105-1125.

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature neuroscience*, *19*(4), 613-622.

Huneman, P. (2010). Topological explanations and robustness in biological sciences. *Synthese*, *177*(2), 213-245.

Illari, P. M., & Williamson, J. (2012). What is a mechanism? Thinking about mechanisms across the sciences. *European Journal for Philosophy of Science*, *2*(1), 119-135.

Jazayeri, M., & Afraz, A. (2017). Navigating the neural space in search of the neural code. *Neuron*, *93*(5), 1003-1014.

Kaiser, M. I., & Krickel, B. (2017). The metaphysics of constitutive mechanistic phenomena. *The British Journal for the Philosophy of Science*.

Kaplan, D. M. (2011). Explanation and description in computational neuroscience. *Synthese*, *183*(3), 339-373.

Kaplan, D. M. (2015). Moving parts: the natural alliance between dynamical and mechanistic modeling approaches. *Biology & Philosophy*, *30*(6), 757-786.

Kaplan, D. M., & Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of science*, *78*(4), 601-627.

Kästner, L. (2018). Integrating mechanistic explanations through epistemic perspectives. *Studies in History and Philosophy of Science Part A*, *68*, 68-79.

Keyser, S. J., Miller, G. A., and Walker, E. (1978). *Cognitive science, 1978*. An unpublished report submitted to the Alfred P. Sloan Foundation, New York.

Kirkpatrick, K. L. (2022). Biological Computation: Hearts and Flytraps. *Journal of Biological Physics*, 48 (1): 55-78.

Krakauer, J. W., Ghazanfar, A. A., Gomez-Marin, A., MacIver, M. A., & Poeppel, D. (2017). Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, *93*(3), 480-490.

Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature neuroscience*, *21*(9), 1148-1160.

Laakso, A., & Cottrell, G. (2000). Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical psychology*, *13*(1), 47-76.

Lee, J., & Dewhurst, J. (2021). The mechanistic stance. *European Journal for Philosophy of Science*, *11*(1), 1-21.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Levy, A., & Bechtel, W. (2013). Abstraction and the organization of mechanisms. *Philosophy of science*, *80*(2), 241-261.

Machamer, P., Darden, L., & Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of science*, *67*(1), 1-25.

Maley, C. (2022). Medium Independence and the Failure of the Mechanistic Account of Computation. *Ergo*.

Marr, D. (1982). *Vision*. Freeman and Company, San Francisco.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *200*(1140), 269-294.

McCaffrey, J. (forthcoming). Evolving Concepts of Functional Localization. *Philosophy Compass*.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.

Meyer, R. (2020). The non-mechanistic option: Defending dynamical explanations. *The British Journal for the Philosophy of Science*.

Milkowski, M. (2013). *Explaining the computational mind*. MIT Press.

Miłkowski, M. (2016). Explanatory completeness and idealization in large brain simulations: A mechanistic perspective. *Synthese*, *193*(5), 1457-1478.

Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., ... & Kay, K. (2018). Cognitive computational neuroscience: a new conference for an emerging discipline. *Trends in cognitive sciences*, *22*(5), 365-367.

Newell, A., & Simon, H. A. (1972). *Human problem solving* (Vol. 104, No. 9). Englewood Cliffs, NJ: Prentice-hall.

Niv, Y. (2021). The primacy of behavioral research for understanding the brain. *Behavioral Neuroscience*.

Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., & Semenuks, A. (2019). What happened to cognitive science?. *Nature human behaviour*, *3*(8), 782-791.

Palmeri, T. J., Love, B. C., & Turner, B. M. (2017). Model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 59-64.

Pessoa, L. (2022). *The entangled brain: How perception, cognition, and emotion are woven together*. MIT Press.

Piccinini, G. (2004). The First computational theory of mind and brain: a close look at McCulloch and Pitts's "logical calculus of ideas immanent in nervous activity". *Synthese*, *141*(2), 175-215.

Piccinini, G. (2007). Computing mechanisms. *Philosophy of Science*, *74*(4), 501-526.

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford: Oxford University Press.

Piccinini, G. (2020). Neurocognitive Mechanisms: Explaining Biological Cognition. Oxford: Oxford University Press.

Piccinini, G. (2022). "Neurocognitive Mechanisms: Some Clarifications," *Journal of Consciousness Studies*, 29.7-8 (2022), pp. 226-250.

Piccinini, G., & Craver, C. (2011). Integrating psychology and neuroscience: Functional analyses as mechanism sketches. *Synthese*, *183*(3), 283-311.

Piccinini, G., & Shagrir, O. (2014). Foundations of computational neuroscience. *Current opinion in neurobiology*, *25*, 25-30.

Pinto, N., Cox, D. D., & DiCarlo, J. J. (2008). Why is real-world visual object recognition hard? *PLoS computational biology*, *4*(1), e27.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Reviews Neuroscience*, *1*(2), 125-132.

Povich, M. (2015). Mechanisms and model-based functional magnetic resonance imaging. *Philosophy of Science*, *82*(5), 1035-1046.

Pylyshyn, Z. W. (1984). *Computation and cognition*. MIT Press.

Rathkopf, C. (2018). Network representation and complex systems. *Synthese*, *195*(1), 55-78.

Rescorla, M. (2018). An interventionist approach to psychological explanation. *Synthese*, *195*(5), 1909-1940.

Riesenhuber, M., & Poggio, T. (2000). Models of object recognition. *Nature neuroscience*, *3*(11), 1199-1204.

Ritchie, J. B. (2019). The content of Marr's information-processing framework. *Philosophical Psychology*, *32*(7), 1078-1099.

Ritchie, J. B., & Carlson, T. A. (2016). Neural decoding and "inner" psychophysics: A distance-to-bound approach for linking mind, brain, and behavior. *Frontiers in neuroscience*, *10*, 190.

Ritchie, J. B., & Piccinini, G. (2018). Computational implementation. In *The Routledge handbook of the computational mind* (pp. 192-204). Routledge.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, *65*(6), 386.

Ross, L. N. (2015). Dynamical models and explanation in neuroscience. *Philosophy of Science*, *82*(1), 32-54.

Ross, L. N. (2018). Causal selection and the pathway concept. *Philosophy of Science*, *85*(4), 551-572.

Ross, L. N. (2021). Causal concepts in biology: How pathways differ from mechanisms and why it matters. *The British Journal for the Philosophy of Science*.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing*: *Vol. 1*. Cambridge, MA: MIT Press/Bradford Books.

Rust, N. C., & Stocker, A. A. (2010). Ambiguity and invariance: two fundamental challenges for visual processing. *Current opinion in neurobiology*, *20*(3), 382-388.

Salmon, W. C. (1984). *Scientific explanation and the causal structure of the world*. Princeton University Press.

Piccinini, G., & Scarantino, A. (2011). Information processing, computation, and cognition. *Journal of biological physics*, *37*(1), 1-38.

Sejnowski, T. J., Koch, C., & Churchland, P. S. (1988). Computational neuroscience. *Science*, *241*(4871), 1299-1306.

Serre, T. (2019). Deep learning: the good, the bad, and the ugly. *Annual review of vision science*, *5*(1), 399-426.

Shagrir, O. (2010). Marr on computational-level theories. *Philosophy of science*, *77*(4), 477-500.

Shagrir, O. (2022). *The Nature of Physical Computation*. Oxford: Oxford University Press.

Shapiro, L. A. (2017). Mechanism or bust? Explanation in psychology. *The British Journal for the Philosophy of Science*.

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind & Language*, *22*(3), 246-269.

Turner, B. M., Forstmann, B. U., Love, B. C., Palmeri, T. J., & Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. *Journal of Mathematical Psychology*, *76*, 65-79.

Weiskopf, D. A. (2011). Models and mechanisms in psychological explanation. *Synthese*, *183*(3), 313-338.

Weiskopf, D. A. (2017). The explanatory autonomy of cognitive models. *Explanation and integration in mind and brain science*, 44-69.

Williams, D., & Colling, L. (2018). From symbols to icons: The return of resemblance in the cognitive neuroscience revolution. *Synthese*, *195*(5), 1941-1967.

Zednik, C. (2017). Mechanisms in cognitive science. In *The Routledge handbook of mechanisms and mechanical philosophy* (pp. 389-400). Routledge.

Zednik, C. (2019). Models and mechanisms in network neuroscience. *Philosophical Psychology*, *32*(1), 23-51.