How etiology does, and how it should, shape modeling choices in neuroscience

Lotem Elber-Dorozko

Abstract: It is common today in machine-learning research for scientists to design and train models to perform cognitive capacities, such as object classification, reinforcement learning, navigation, language processing and more. Neuroscientists compare the processes of these models with neuronal activity, with the purpose of learning about computations in the brain. These machine-learning models are constrained only by the task they must perform. Therefore, it is an interesting, somewhat surprising, scientific finding that the workings of these models correlate with neuronal activity, as several prominent papers reported. Such correlations are usually explained by suggesting that the model and the brain have learned or adapted to perform similar tasks. This general approach has much promise. However, I argue that to the extent that its aim is to explain how cognitive capacities are performed in the brain, it must be much more careful in making etiological claims. For not every function performed by brains is usefully treated as a function that is the result of a distinct optimizing process. Being too permissive about such functions, and referring to adaptation too quickly, may lead to the acceptance of models that are very poor descriptions of cognitive capacities.

1. Introduction

As the capabilities of machine-learning algorithms grow, it is becoming increasingly common in the cognitive sciences to utilize the following methodology: identify some cognitive capacity,[1] use machine learning research to build and train algorithms to achieve this capacity, and compare the workings of these algorithms with neuronal activity. When neuronal activity is found to correlate with processes in the machine-learning algorithm, this finding is worthwhile for two reasons - First, we gain a new way to predict neuronal activity, often with better accuracy than previous models. Second, the finding of correlation suggests that computation in the brain is similar in some ways to the machine-learning algorithm. Such work was done for object recognition (Cao and Yamins 2022a; Yamins et al. 2014; Yamins and DiCarlo 2016; Zhuang et al. 2021), reinforcement-learning (Cross et al. 2021), language processing (Goldstein et al. 2022; Kell et al. 2018; Schrimpf et al. 2021), navigation[2] (Banino et al. 2018; Cueva and Wei 2018), orientation during self-motion (Mineault et al. 2021) and more.

Borrowing from (Yamins et al. 2014), I will henceforth call this methodology 'performance-based methodology', because it aims to create models that can perform capacities that people perform. This methodology is not unlike Marr's (1982) framework for levels of analysis: it begins by describing the performed computation, then it identifies an algorithm that can perform this computation, and finally it searches for the algorithm's neuronal correlates. This approach emphasizes the

[1] In lieu of a better term, I use the term 'capacity' in this paper to indicate some behavior a system can perform. Under this meaning one can say that people have the capacity to bite their nails or fidget. Some may interpret the word to indicate some relation to fitness, this is not how I use it here. In fact, the relation between investigated capacities and fitness is one of the questions I discuss throughout the paper.

[2] Navigation is a slightly different case because neuronal activity is already characterized as representing location in a grid like manner, and therefore neuronal activity is well explained with a simple concept. Scientific works show how these representations arise as part of learning navigation-related tasks.

usefulness of environmental and functional constraints in modeling neuronal activity – instead of focusing on biological processes in the brain, this approach investigates the tasks an organism faces and suggests ways in which they can be addressed. At least in the step of constructing the algorithm for the cognitive capacity, this approach also minimizes the importance of physical, developmental, or evolutionary constraints – the only constraint on the algorithm is that it achieves high performance on the relevant tasks. For this reason, it is often a pleasant surprise for scientists when they discover similarities between the model and neuronal activity.

How can such similarities between brain processes and artificial algorithms be explained? Researchers point to shared features in the ways the two capacities come about. One line of argument is that the performed task strongly constrains the algorithms that underly it, so that all systems that can perform a task are very similar. As Kell et al. (2018, 630) write: "The underlying hypothesis was that everyday recognition tasks may impose strong constraints on the auditory system, such that a model optimized to perform such tasks might converge to brain-like representational transformations". Yamins and DiCarlo (2016, 360) write: "within the class of HCNNs [hierarchical convolutional neural networks], there appear to be comparatively few qualitatively distinct, efficiently learnable solutions to high-variation object categorization tasks, and perhaps the brain is forced over evolutionary and developmental timescales to pick such a solution".

Others, point to similarities between adaptation and learning in people, and architecture choice and training in artificial models. Hasson et al. (2020, 425): "Similar to natural selection, the family of models to which both ANNs [artificial neural networks] and BNNs [biological neural networks] belong optimizes parameters according to objective functions to blindly fit the task-relevant structure of the world, without explicitly aiming to learn its underlying generative structure". Cao and Yamins (2022b) write: "Structurally, the kinds of search through possibility space that the modeler undertakes are analogous to the kinds of search that result in competent adult brains shaped by evolutionary and learning and development … perhaps we should not be so surprised that our two known solutions to one of those problems (that of visual object classification) were arrived at by structurally similar routes."

Here, I argue that such a view of the etiological processes leading to cognitive capacities overly emphasizes the similarity between the etiologies of brain and artificial processes, while ignoring crucial differences. Often, this appeal to etiological similarity add to the support of the models as good models of the cognitive capacities, when the differences in etiology should rule out the suggested models as reflecting brain processes.

As has been pointed out before (Novick 2023), generally, biological properties are the result of the interaction of several different elements. These include selection of certain properties to improve fitness, but also structural, developmental and physical constraints biasing and limiting the range of possible properties. Not every cognitive capacity is well-described as the result of an optimizing process, such as adaptation or learning (Gould and Lewontin 1979). Some capacities may be better described as side-effects of the selection and learning of other functions, and other capacities are too specific or too broad to be described as the result of a distinct optimizing process. Using the described 'performance-based methodology' to model behaviors that are not the result of a distinct optimization processes will overlook important constraints on the way these behaviors are performed, and therefore is unlikely to result in processes that are similar to brain processes.

This point does not mean to overemphasize history. It is not the case that two systems differ because they have different histories, as it is certainly possible for a capacity to be performed in the same way in two systems with different histories. Instead, the point here is that, given that scientists are in the middle of a process of identifying how cognitive capacities are performed, just like neuronal and behavioral data, etiological considerations also serve as crucial points of evidence to whether the model can explain how a capacity is performed.

Finally, papers employing ‘performance-based methodologies’ report findings of similarities, be it correlations or causal relations, between artificial and brain processes. This is often taken as strong evidence that artificial neural networks are good models for computation in the brain. However, this paper argues that while identification of correlates or mapping of causal relations between a model and the brain can help us learn about computation in the brain, it is not, in and of itself, decisive evidence for a specific brain computation. Computational models can yield significant correlations when compared with systems that are designed to perform an essentially different computation from the model, as several scientific publications have shown (Elber-Dorozko and Loewenstein 2018; Jonas and Kording 2017; Marom et al. 2009). Generally, we should expect many various computations to correlate with neuronal activity, and therefore considerations of evolutionary and developmental processes cannot be eliminated, even with much empirical data about neuronal activity. Therefore, putting too much weight on correlational data while being lax with etiological considerations, may lead to erroneous attribution of computation to the brain.

The next section describes one example of the methodology this paper addresses – modeling object classification. Section three will describe the argument for the importance of considering other etiological features affecting cognitive capacities in addition to an optimizing process. It will make this argument by demonstrating three ways in which failure to do so can lead us astray. Then, section three will discuss the difficulties this argument raises for current scientific practice and will suggest a way forward. Finally, section four will address some objections to the argument of the paper, most notably the objection that the neuronal data demonstrates that we identified the right computations.

2. An example for performance-based modeling – object classification

The case of object classification is one well-known example for the use of performance-based models to explain neuronal activity. In their famous paper, Yamins et al. (2014) create a model that can perform an object classification task at near human performance; the model can classify objects from images in various perspectives into one of eight categories: animals, boats, cars, faces, etc.

The architecture of the model is inspired by the structure of the visual 'ventral stream' in the brain (the areas associated with object recognition) in that it includes several feedforward 'layers' where the connectivity between layers is determined according to the 'Linear-Non Linear' (LN) view of neuronal processing, so that the function performed by the neurons is some linear operation on neuronal activity in the previous layer (the 'input neurons'), followed by a non-linear operation. However, the model does not aim to copy neuronal processing, only to use it as an inspiration to successfully perform the task.

To achieve high performance on the object classification task the model is 'trained': the model performs a task that is similar to the test task in principle but with different image input and different semantic categories for classification. Then, after each input, the weights of the connections between 'neurons' in the model are slightly changed to decrease the error on that input. Through cumulative change in the weights, the model 'learns' to perform the task well. Finally, Yamins et al. (2014) chose a model architecture that had the best performance from variety of architectures. They found that the chosen model was able to perform object classification at a human level on the test task, which was new to the model. Crucially, the model was only trained to perform the training task as best as possible, and information about neuronal activity was not used during training.

After training the model, Yamins et al. (2014) recorded neuronal activity in visual areas of monkeys and discovered that the activity of simulated neurons in the highest layer in their performance-optimized model was able to predict activity in the inferior temporal cortex (IT). IT is a 'high' area in the ventral stream, which receives inputs after several stages of neuronal processing, and can support object categorization for a variety of object positions over a wide range of tasks. By linearly regressing the activity of IT neurons on the activity of simulated 'neurons' in the highest layer of their model, they were able to predict 48.5 ±1.3% of the variance in activity in individual neurons in IT across the presentation of 1600 different photos. This is a two-fold improvement in prediction over the other, non-performance-optimized, models they tested. Moreover, Yamins et al. (2014) discovered that intermediate layers in their model were able to predict 51.7± 2.3% of the variance of neuronal activity in the intermediate brain area V4, while the first and last layer in the model predicted a much smaller fraction of the variance. Thus, they found strong correlates between neuronal activity and simulated activity in their model, which fitted with the processing stages in the model and in the brain.

Yamins et al. conclude, in a paragraph which emphasized the role of etiological considerations in building models for neuronal computation: "[the paper presents] a top-down perspective characterizing IT as the product of an evolutionary/developmental process that selected for high performance on recognition on tasks like those used in our optimization... This type of explanation is qualitatively different from more traditional approaches that seek explicit descriptions of neural responses in terms of particular geometrical primitives".

In a follow-up paper, they demonstrate how they view machine learning algorithms as models for neuronal processing in the ventral pathway (Fig. 1). They write: "HCNNs are good candidates for models of the ventral visual pathway. By definition, they are image computable, meaning that they generate responses for arbitrary input images; they are also mappable, meaning that they can be naturally identified in a component-wise fashion with observable structures in the ventral pathway; and, when their parameters are chosen correctly, they are predictive..." (Yamins and DiCarlo 2016).

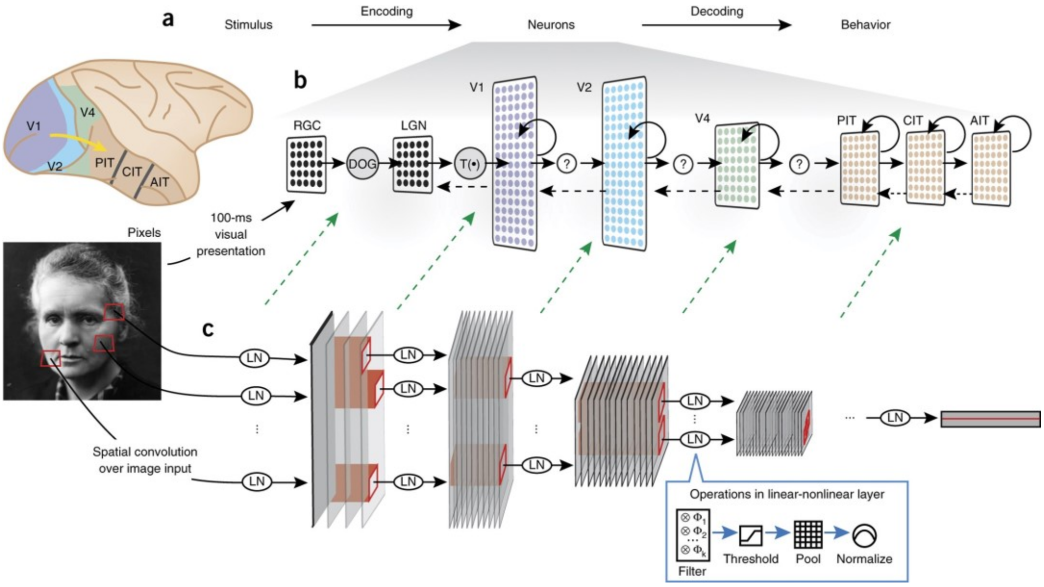

Fig. 1, from (Yamins and DiCarlo 2016), a performance-based model for object classification (**c**) as a model of computation in the brain (**b**). Each layer in the model is mapped to an area in the brain, with corresponding processing stages.

These last two quotes demonstrate two different ways in which the results of the performance-based methodology can be used. First, machine learning algorithms as means to predict neuronal activity is a useful shift from the 'explicit descriptions … in terms of particular geometrical primitives' that Yamins et al. (2014) talk about, because it allows scientists to describe neuronal activity even when it does not resemble a known concept. A second, stronger, claim scientists can make based on results of these correlations go beyond description of neuronal activity to argue that these results are evidence that the performance-based model can answer the question of how the brain performs the relevant cognitive capacity. As (Yamins and DiCarlo 2016) write: "HCNNs are good candidates for models of the ventral visual pathway". This paper targets the second, stronger, claim about computation in the brain, which is often explicitly stated, or otherwise may be tacitly implied.

Performance-based methods have been used to predict neuronal activity for a variety of capacities, including reinforcement-learning (Cross et al. 2021) - where brain activity of participants playing video games was found to correlate with activity in deep layers of a model that was trained to play the same games from inputs of images to outputs of actions; and language processing (Goldstein et al. 2022) – where neuronal activity while listening to a podcast could be predicted from representations created by language models, to name a few.

In some cases, it has been explicitly argued that performance-based models are models for neuronal computation. (Goldstein et al. 2022) write: "[T]he human brain and autoregressive DLMs [deep language models] share three fundamental computational principles"; (Zhuang et al. 2021) claim: "[These results] present a strong candidate for a biologically plausible computational theory of primate sensory learning." It has even been suggested that such computational models whose simulated activity maps onto neuronal activity according to specific criteria, met by the model in Yamins et al. (2014), are mechanistic explanations of how the brain performs the capacity (Cao and Yamins 2022a).[3]

Following impressive results of correlations between neuronal activity and performance-based models from a variety of papers (Banino et al. 2018; Cross et al. 2021; Cueva and Wei 2018; Mineault et al. 2021; Schrimpf et al. 2021; Yamins et al. 2014) it may seem that this methodology can yield new understanding of the underlying computation for any capacity of our choosing. However, in the next section I point out that for many capacities, some of them investigated in the cited papers, portraying them as the result of a distinct optimizing process misses crucial aspects of how they came about. If one ignores these aspects, although simulated activity may show some mapping to neuronal activity, the computational models are likely to be different in important ways from the ones employed by the brain. For they will miss important constraints on how the capacity is performed at present.

3. How etiological considerations matter

The next three sub-sections describe three different ways in which the etiology of a capacity majorly affects how it is currently performed. Specifically, a capacity that is the result of a distinct optimizing process is performed in a substantially different way from a capacity that had additional constraints in its history. Thus, models that erroneously treat a cognitive capacity as the former will substantially diverge from the true processes underlying the capacity.

A. Modeling side effects, rather than adapted functions

[3] See (Craver 2007; Kaplan and Craver 2011; Piccinini 2015) for detailed frameworks of mechanistic explanations

The brain does many things, some of them it has been adapted to do and some are ‘side-effects’ of other evolutionary or developmental processes. Biological functions have been extensively discussed in philosophy (Boorse 2002; Cummins 1975; Millikan 1989; Neander 1991; Wouters 2005). The major question has been what differentiates the *functions* of the system from other things the system does. To give the oft used example, the heart both pumps blood and makes thumping sounds, but we usually only take the former to be its function. On perspectivalist views of function, the functions of the system are not an objective matter, but rather depend on the interests of the observers (Craver 2013). On such views the heart’s function may well be to make thumping sounds if the observer is interested in building stethoscopes. This observer may also be interested in explaining the underlying mechanism that is responsible for the thumping sounds.

Another set of views take functions to be an objective matter. One such popular view of functions is the ‘selected-effects’ view. This view describes functions by reference to their evolutionary history; the function of a system is to bring about effects that in the past were relevant to its selection (Millikan 1989; Neander 1991). Hence, hearts have the function of pumping blood but not the function of making thumping sounds, because only their ability to pump blood was causally relevant to the existence of the organism today. Therefore, there is a difference between functions the system has because they were previously relevant for its selection, and functions the system can perform, i.e., side-effects.

It is not my intention to make an argument in favor of one view or other of function. Nonetheless, the distinction between ‘side-effects’ and ‘selected-effects’ is relevant to the epistemological practice of building and assessing computational models for cognitive tasks. This, because capacities that are considered side-effects according to the selected-effects view are, by definition, very unlikely to be the result of a process optimizing them as capacities. Instead, they are the result of processes optimizing other, related, capacities. Therefore, using performance-based methodology, they are unlikely to be given a computational model that is similar to the computation that takes place in the brain.

As a thought experiment, consider a scientist who encounters for the first time a lightbulb. The scientist has no idea what the function of the light bulb is, or if it even has one. She notices that the lightbulb emits heat and tries to explain how it does so. She comes up with a model for a heat emitting device – a radiator. The radiator is just as good at emitting heat as the lightbulb. Therefore, according to the performance-based methodology it is a model that can be compared with the activity of the light bulb. Furthermore, correlations between the activities of the two may even be identified, as I also argue in section 4. For example, both heat up when connected to electricity. Nonetheless, there is some deep sense in which the scientist missed how the lightbulb emits heat – it does so via a mechanism that was designed to emit light. The lightbulb emits heat, but it *has* the function of emitting light, and this puts specific constraints on its mechanism for emitting heat, leading to electricity being passed through thin cords in a glass tube, for example. Models constructed specifically to emit heat are likely to miss these constraints. Similar scenarios will occur if someone tries to explain how a coffee machine emits such a strong noise using a performance-based approach; the issue isn't that the suggested models do not make coffee, but rather that they are very unlikely to suggest the right answers for the source of the noise – grinding coffee beans and foaming milk. Therefore, they are very unlikely to come up with a model that is similar to how the coffee machine produces noise.

In relation to human cognition, we can consider chess playing. The performance-based methodology would build a machine-learning algorithm that can play chess and compare its activity with brain activity. Such chess-playing models have already been created and rivaled human champions. However, according to the selected-effects view, people *can* play chess, but they do not *have* the function of playing chess. Brains were not adapted for chess playing so it would be astonishing to discover that neuronal computation is similar to algorithms designed specifically for chess playing, such as deep blue (Campbell, Hoane, and Hsu 2002). An accurate computational model of human chess playing will take into account that this capacity utilizes mechanisms that were adapted for other purposes. Similar points can be made regarding driving, baking a cake and synchronized swimming. While some of these capacities are useful today, such as driving, it is clear that they exist not as the result of a dedicated optimizing process, but as a side-effect of our perceptual and motor abilities more generally.

One could argue that, while some capacities are clearly not the result of a distinct adaptive process, they may be the result of another optimizing process, namely learning. Therefore, they will be well-modeled as such. It is still an open question the extent to which the learning of a specific skill can be treated as independent of other brain processes. Generally, we would expect the learning of new skills to rely at least on existing perceptual and motor capacities. Whatever one's view on this question, it should figure in the discussion when modeling learned skills. Moreover, we should be careful not to overestimate the ability of learning to shape brain processes in idiosyncratic experimental tasks. Thus, building models that are optimized to perform 'side-effects' or learned tasks without considering additional constraints on the optimization process is unlikely to yield similar processes to how they are performed in the brain.

B. Modeling overly specific capacities

Not every capacity that is considered a function according to the selected-effects view will be the result of a specific optimizing process. Some cognitive functions may only be partial descriptions of the capacities that brains have adapted to have. When considering capacities that are the result of evolution, it is useful to consider what evolutionary psychologists call 'Darwinian modules' (not to be confused with Fodor's modules, which are characterized differently) – capacities that are the result of a distinct evolutionary process (Machery 2007b, 2007a). As Machery (2007b) writes: "evolutionary psychologists are adamant that many competences, such as reading, programming in C++, and piloting an airbus, are not underwritten by dedicated modules. There is no module whose evolved function is, say, to read, since, obviously, reading is a recent cultural invention. Rather, reading is underwritten by a collection of modules that evolved for other reasons."

The examples given by Machery may be considered side-effects from an evolutionary perspective, since they are too recent to be the result of an adaptive process. But the general point serves well for capacities that are not side-effects, but simply partial descriptions of capacities with distinct optimization processes.

The fact that our ancestors were able to distinguish zebras and tigers increased their fitness, and this capacity would be considered a function according to the selected-effects view, but we do not think ancestral brains have adapted for this specific task independently of other perceptual tasks, and so we do not expect the capacity to differentiate zebras and tigers to be a 'Darwinian module'. Imagine a scientist using the performance-based methodology to explain how people distinguish between zebras and tigers. They build a model and train it to make this distinction. With current technology, the model will probably do very well. But this model is likely to solve this problem in a very different manner than people. It may classify black and white objects as zebras and the rest as tigers, for example. People are unlikely to use this method because they don't learn to distinguish zebras and tigers alone, they must learn perform a much more complex capacity, also to distinguish zebras from cats, chess boards, and cross walks. Thus, because the scientist chose an overly narrow description of a capacity as the target of a distinct optimization process, it is very unlikely they come up with good models.

As another example, consider an attempt to explain how people swim. The performance-based approach will attempt to come up with the best model for controlling movement in water. This model will likely resemble a fish. It will widely diverge from how people use their bodies to swim, because it ignores other constraints on the human body, specifically that it needs to be able to also move on land, and therefore cannot be distinctly optimized for swimming.

The point made here is not unlike that famously made by (Gould and Lewontin 1979). They criticize evolutionary biology as prone to telling 'just-so' stories, describing any biological feature as an adaptation without supporting evidence. Here I suggest an even stronger interpretation of this point. Not only is it not obvious for many capacities that they are adaptations, but for some, such as baking or playing computer games, it is obvious that they cannot be reasonably described as the result of a unique optimization process, without further constraints. This insight can be useful in modeling cognitive capacities.

C. Modeling tasks with unnatural data

There is a long line of researchers advocating for examining behavior in more natural scenarios (Gibson 1979; Krakauer et al. 2017; to name a few). This section will do the same, albeit for different reasons. For even if one is convinced that conducting a simple experiment in the lab is a good proxy for behavior in natural environments, this does not mean that training models on the same unnatural stimuli will yield models that are similar to the computation in the brain.

There is some overlap between this subsection and the previous two subsections, in that unnatural stimuli tend to be either stimuli that people do not encounter in their natural environment and therefore we do not expect them to have adapted to perform the corresponding task, or they are stimuli that are overly simplified and do not capture the true complexity of the modeled capacity. In both cases training models on such stimuli is unlikely to lead to models that resemble computation in the brain.

Consider as one example a two-armed bandit task where participants repeatedly choose between two actions (Fox et al. 2020). Presumably people use in this case a general system for decision making, which can be utilized in various scenarios, with 3, 4, or infinite options, where states and actions may change in unpredictable ways, etc. Nonetheless, having people perform this task in the lab may lead to worthwhile, albeit simplified, insights (Fox et al. 2020; Shteingart, Neiman, and Loewenstein 2013). However, if we train a model in our simple scenario there is no promise that it will be able generalize to other cases and in this sense, it will significantly differ from computation in the brain. The only exception is if we think that the brain went through a specific optimization process to repeatedly decide between two options. Thus, training models on experimental tasks that are used to assess human behavior is likely to lead to models performing capacities that fall into one of the discussed pitfalls –

either they are unnatural ‘side-effects’, or they are overly simplified. Both these options describes many capacities modeled in scientific practice.

D. Current scientific practice

While current neuroscientific practice strives to use stimuli and tasks that are as natural and complex as possible, in various cases it is still quite clear that the capacities that models are trained to perform cannot be capacities that are the result of a distinct optimization process.

Consider object classification. Clearly, identifying objects is beneficial for survival. However, when delving into the details we see that the model was trained to classify a restricted set of objects from a variety of photos where objects are placed on unmatching backgrounds (see Fig. 2).

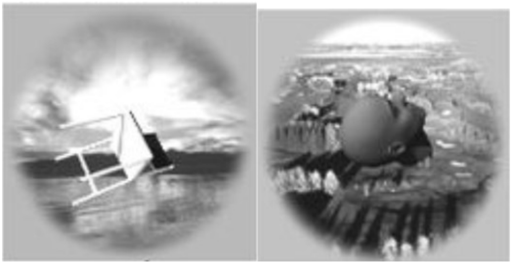

Fig. 2. Example of two test images from (Yamins et al. 2014). Left – a chair. Right - a face.

This choice for training data is understandable, as matching backgrounds may lead the model to use the background to classify the object. Nonetheless, the result is that the task is clearly one for which there was no distinct optimization process. The ability to perform this specific task is a side-effect of classifying objects in natural environments, in which objects are placed in specific contexts in time and in space, and so we would expect the computation performed by the brain in this task to be different from the trained model in (Yamins et al. 2014). Moreover, even in cases where training is done on natural images, as in (Zhuang et al. 2021) there is room to wonder if this task should be considered the result of a distinct optimization process. Perhaps it is more reasonable to say that object classification adapted and was learned for actively extracting relevant information from moving visual scenes, in a specific environmental context, into a wide and complex array of categories. Moreover, proponents of embodied cognition have suggested that it is likely that perception has adapted to support actions that contribute to fitness rather than to accurately represent the environment (Proffitt 2006) and (Bowers et al. 2023) have suggested other selection pressures on the ventral visual stream besides maximizing classification accuracy.

Similar claims can be made regarding other cognitive capacities. For example, models for reinforcement learning from visual inputs are generally trained and tested in video game environments (Cross et al. 2021). While these environments are meant to imitate decision making processes, they substantially differ from natural environments in various elements, including a simple and discrete structure of states and actions, and explicit rewards. Therefore, one could call the ability to play video games a side-effect of the human capacity for decision-making. To the extent that playing video games relies on a decision-making capacity, it is a capacity for simpler environments than natural environments, one which is unlikely to be the result of a distinct optimization process in the human brain that deals with complex, changing, and continuous environments.

The claim that current scientific practice ignores important constraints on cognitive capacities that are the result of specific etiological processes is not meant to discourage this specific area of research. It is certainly an area worth pursuing, in which scientists are demonstrating how machine-learning models can perform more complex and impressive capacities. What this paper aims to do instead is to point out a specific relevant domain which scientists should pay attention to when assessing such models; it is not enough that a model can perform the capacity and that there are correlates between the model and neuronal activity. An artificial neural network that has come to perform some capacity through a distinct optimization process can serve as a good model for that capacity to the extent that the cognitive capacity has also come about through a distinct optimization process.

E. Identifying capacities that result from distinct optimization processes

So far, this paper has described several cases where, intuitively, the modeled capacity is not a capacity for which there was a distinct optimization process and therefore we do not expect the model to correspond to computation in the brain. The question that arises is how one can identify capacities that can be well-modeled using the performance-based methodology.

One aspect of this issue has been described as the 'grain problem' – evolutionary pressures can be described at finer or at courser grain and there doesn't seem to be a principled reason to choose one grain level over the other. As (Atkinson and Wheeler 2004) point out, this is also true for phenotypic traits, so one cannot appeal to them to decide on the right grain level. There clearly isn't a clear-cut answer to the question of which biological capacities can be usefully modeled as the result of distinct processes, because all capacities in an organism depend on each other.

Although this issue stands, it is an empirical fact that we can divide many biological systems into subsystems in a way that aids the explanation of phenomena – bodies can be divided into organs and sub-systems such as the immune and endocrine systems, organs into cells, cells divided into organelles, and so forth. In physiology as well, while it is a scientific achievement to say which processes can be modeled as distinct functions, some processes are clearly poor contenders for performance-based modeling. Consider a model for the immune system that can only account for its response to a limited set of pre-defined pathogens. While this model may explain part of the immune system's function, it is a very poor model for the immune system as a whole, because it misses a crucial aspect of its function, specifically that it must be able to respond to new pathogens.

Dividing cognition into systems that can be analyzed distinctly is a much more challenging task. There are many debates about cognitive ontology and whether cognitive capacities can be individuated in space or whether they can be treated as consistent in time (McCaffrey 2023). Furthermore, there are debates about whether cognition can be divided into Darwinian modules at all (Machery 2007b; Quartz 2002).

In dealing with these challenges a few notes are worth considering. First, it may turn out that good models for some cognitive capacities must be a very general ones, and that perception cannot be modeled independently of decision making, for example. Although these results mean greater challenges to scientists, contending with these issues may be necessary to explain cognition. Nonetheless, it is worth mentioning that for some cognitive capacities, there have been suggested models that focus specifically on the performance of the capacity, and those have been well-supported and widely accepted as good models.

It is known that, in certain birds, the neurons in the nucleus magnocellularis and the nucleus laminaris (areas in the brain stem) serve as a system that implements the 'Jeffress model' to compute the difference in time delay of sound between the two ears (Ashida and Carr 2011). This computation is the basis for sound localization for certain frequencies. There is also strong evidence that an area in the central complex of the fly brain implements a computational model known as the 'ring attractor' to represent head direction (Turner-Evans et al. 2020). Hence, it is not impossible to model some cognitive capacities with 'performance-based models'[4]

How can we identify capacities that are fit to performance-based modeling? I suggest that to identify computation in the brain we should appeal to an interplay of evidence with common-sense assumptions from different domains. In this interplay, etiological considerations should be used to challenge existing models and to suggest ways such models can be made more plausible, rather than being abused post-hoc to support existing models. Attempting to model capacities that are more plausibly the result of a distinct optimizing process is likely to lead to better prediction of behavior and neuronal activity. Similarly, correlates with neuronal activity is also evidence for the etiology of capacity. If a 'performance-based' computational model could predict 99% of neuronal variance (which is currently not likely due to individual heterogeneity, [Cao and Yamins 2022a]), this would be strong evidence also for the etiology of the capacity – this is the computation the brain has historically been specifically optimized to perform. Additionally, anatomical evidence about sub-structures and experiments testing for double dissociation can also be used and have been used to provide evidence for distinct functions. However, etiological considerations never disappear entirely, even a 99% prediction of neuronal activity would not convince us that the brain is computing the location of Mars relative to Neptune (see also section 4B). Thus, there is reciprocity between considerations of

[4] While the Jeffress model and the ring model are not deep learning models and have not been optimized through gradual change to perform the capacity, they can still be considered 'performance-based models' in the sense that they can perform the capacity and were built for this purpose, without knowledge of brain processes.

etiology and considerations of similarity to neuronal activity,[5] and one should be careful not to put too much weight on the latter.

It is not impossible to suggest the right computational model without etiological considerations. As mentioned before, the etiology of a capacity does not define it, and two systems can perform a capacity in the same manner despite having very different etiologies. But from an epistemological perspective, when attempting to identify the computations and mechanisms underlying capacities, using models that ignore important constraints on how a capacity came about is very likely to be unfruitful.

In the next section I present some objections to this paper's argument.

4. Some objections

A. Neuronal correlates can fully support specific computations

One evident objection to the claim that computational must consider etiology more carefully, is to note that this argument completely ignores the successes of this practice. The described scientific projects in previous sections identified correlations between neuronal activity and simulated activity in the model. Is this not evidence that these are the models that are implemented in the brain?

Although this claim seems obviously true, several scientific publications have demonstrated that it is entirely possible to identify correlations and causal relations that map with one computational model, when the system is designed to perform a completely different computation[6] (Elber-Dorozko and Loewenstein 2018; Jonas and

[5] Interestingly, (Atkinson and Wheeler 2004) suggest a similar approach: "Ideally there is a dynamic and mutually constraining relationship between attempts to infer architectural solutions from adaptive problems and attempts to infer adaptive problems from architectural solutions."

[6] Some readers with a philosophical background may be reminded of 'the triviality arguments about computational implemental' (Sprevak 2018). These arguments expose why it is problematic to define computation as mapping between a physical system and a computational model. It has been argued that, without constraints on the mapping relations between the physical system and the computation, any physical system can be mapped to any computational model. If computation

Kording 2017; Marom et al. 2009). Famously, Jonas and Kording (2017) utilized standard neuroscientific methods to understand the workings of a microprocessor that performed a simple task of booting one of three video games. They arrived at ridiculous results such as a "Donkey Kong transistor or a Space Invaders transistor." – transistors that are taken to have a function that relates only to one specific game, when it is well-known that this is not how microprocessors are designed.

Elber-Dorozko and Loewenstein (2018) analyzed the case of 'action-value representations'. Many previous scientific findings reported brain representation of a variable called 'action-value'. Elber-Dorozko and Loewenstein (2018) specifically designed a model for decision making which does not include any implicit or explicit representation of 'action-value', and demonstrated that standard analyses performed on this model still erroneously identified significant representation of 'action-value'.

These results demonstrate that correlation cannot distinguish competing hypotheses about computation (and, for the same reason, neither do mapping of causal relations). It is easier to understand why this is so when we consider that when performing a correlation analysis, the null hypothesis is that the neuronal activity is *completely* orthogonal to the computational variable. Any other case with sufficient data will result in a significant correlation. Thus, identification of a correlation between neuronal activity and some variable is not an indication that this variable is computed, but only that neuronal activity is not completely orthogonal to this variable (see also (Elber-Dorozko and Loewenstein 2023) for a more detailed argument). Given that any computational variable that performs some capacity is likely to correlate with properties of the inputs and the outputs of the capacity, there are many possible

depends solely on mapping, the resulting picture is one of pan-computationalism. The scientific papers here similarly demonstrate the problem of relying on mapping to identify (rather than define) computation, even within the methods employed in neuroscience.

computational models that correlate with neuronal activity without being identical to neuronal computation.

Even though scientific results of correlation with neuronal activity cannot conclusively support a specific computational model over its competitors, still much can be learned from them. First, if they are not taken as the sole relevant evidence, they can be invaluable in comparing suggested models. Schrimpf et al. (2020) built a platform for comparison of various computational models with neuronal data in a variety of visual tasks. Such comparisons can certainly assist in determining what computational properties lead to closer resemblance to neuronal processing (but see (Bowers et al. 2023) on the domains in which such evidence should be sought). Relatedly, as I argued in the previous section, evidence that neuronal activity correlated with a computational model can also support the hypothesis that the modeled capacity has been distinctly optimized for. But, if it is implausible that the modeled capacity is the result of a distinct optimization process, the evidence from neuronal activity should be overwhelming to convince us that our plausibility considerations have been wrong. So far, very rarely is evidence for neuronal correlates of a model overwhelming.

B. Etiology does not determine computation

The reader may have noticed that the argument made in section 3 moved quickly when discussing what is the ‘right’ and what is the ‘wrong’ computational model for the computation performed in a system. The examples in section 3 and the scientific papers described in 4A, refer either to adaptation or to design as the determinant of the computation the system performs; that is, conclusions about computation are erroneous because they do not fit with what the system was designed to do or with our intuitions about what the system has adapted to do. There are no ‘donkey-kong’ transistors because no transistors were designed as such, and there is no chess playing module because the brain has not adapted or developed for chess-playing. This notion fits with the philosophical view that the question of what a physical system computes depends on its etiology. One could adopt such a view if one takes computing systems to be systems that have the function to perform some computation and this function is defined according to the etiology of the system.

One could, of course, deny that etiological considerations are relevant to determine what the brain computes. There are several popular views of computation in accordance with such claims. Shagrir (2022) argues that the individuation of a computation depends on its semantic content (this would be a non-etiological view only if we take semantic content to not be determined etiologically). Piccinini’s (2015) framework of physical computation describes computing systems as mechanisms that have the function of performing a specific computation. He is explicit, however, that the functions he refers to are not defined by their evolutionary history, but rather by their current causal contributions (2015, chap. 6).

Such views are worthwhile alternatives to the etiological view. Moreover, there are several criticisms of etiological views of function. One such criticism is that it defies our common-sense view of computation to think that the history of a system should be relevant to determine what a system is currently computing (Craver 2013; Piccinini 2015).

As an answer to this criticism, note that these philosophical debates center on the question of what computation *is*, while I aim to discuss questions about how computation can be identified. Thus, one can adopt an ontological view where etiology does not matter for computation, but still agree that it is relevant epistemically. Claims that etiology is irrelevant to current computation often rely on rare cases where etiology and computation come apart. The case of a swamp-person miraculously created de-novo, or the case of a major first mutation which turns out to be beneficial. While it may certainly be true that in these cases what the system computes comes apart from its history, they are too rare to merit overlooking history in general. For swamp-people never happen, and first mutations tend to be small and to build upon previous states. Thus, ontologically it may be true that etiological considerations do not determine what a system computes. However, empirically, for practically all systems we view as computing, their etiology is useful for understanding how they perform the computations: Organisms have evolutionary histories and computers are designed.

Another challenge to the claim that careful etiological considerations are essential is that it is very difficult to know the etiology of various capacities, so it is difficult to see how they can be taken into account in identifying computation, and nonetheless scientists move forward with assigning functions and computations. One answer to this is that scientists do take etiology into account when modeling cognition, explicitly and implicitly. Explicit considerations can be seen in all the citations in this paper referencing optimizing processes as leading causes to cognitive capacities. Implicit considerations are seen in the capacities scientists choose to model. Visual perception is an extremely popular choice, while baking is not.

The point of this paper is that the use of these etiological considerations can be made better by being more selective about the capacities that are treated as the result of a distinct optimizing process, and taking into account other constraints that affect capacities other than optimizing processes. Some such constraints which are available to use without conducting phylogenetic research (which is also important) is to ask ‘what other task properties has the brain adapted/learned to have?’ and ‘what is the relationship between the task learned by artificial models and the task performed by people in natural environments?’. Moreover, as described in section 3, models that have an etiology similar to that of the modeled capacity are also more likely to yield similar behavior and internal processes as brains, leading to potential reciprocal improvement of the models.

Finally, I present a challenge to non-etiological views. it is not clear what epistemological alternative non-etiological views of computation suggest. Without constraints on mapping between computational and physical states an incredibly large variety of computations can be considered to be implemented in a system, as demonstrated in the ‘triviality arguments about computational implementation’ (Sprevak 2018). Therefore, views that deny that etiological considerations are relevant for computation, describe other constraints on the computations implemented in a system. The challenge to these views is to explicate the implications of these constraints to scientific practice. Without such implications to neuroscientific practice, although individuation of computation may be well-defined ontologically, it is not clear how questions about what a system computes can be answered. Etiological considerations at least offer some way to advance in this regard for the vast majority of computing systems.

C. Current models are a good approximation of cognitive capacities

There is worry that my criticism of the functions that neuroscientists model is too harsh. Surely, they are limited, but they are still a great improvement relative to earlier, simpler models and they are making effortful attempts to be realistic. To this I answer that this paper does not aim to invalidate the progress that is achieved with this practice, these performance-based models are certainly a step forward towards more accurate models of cognitive capacities. However, I do suggest that as they are, they cannot yet provide a realistic model for these capacities, and it is useful to keep this in mind. Moreover, claims about similarities in etiological processes between artificial and brain processes to ignore the important differences between them may cause false convictions that such models explain the performance of cognitive capacities from all perspectives.

If neuroscientists wish to claim that their models aim to capture a specific capacity which was created by distinct optimization processes, it would be beneficial if they would do so explicitly. To illustrate, Yamins et al. (2014) may claim that their model describes the first forward pass in the ventral stream where only feedforward connections are relevant and an object is recognized quickly from a single snapshot. This is different from arguing that their model is a model of 'the ventral steam' and may be much more plausible. Then, the question shifts whether it is reasonable that this quick classification in the forward pass is the result of a distinct optimization process.

Finally, to the extent that the computational models created in the performance-based methodology are close to computation in the brain, if one is convinced by the argument in this paper, then it paints a path forward for existing models; rather than aiming to account for more neuronal variance, or improve performance on pre-existing tasks, we should focus on trying to model capacities that were distinctly optimized. While we should not undervalue our current successes, it is important to keep in mind the way still to go.

D. Even when ignoring considerations of adaptation, we may still identify the right computation

Cao and Yamins (2022b) write: "…given a challenging task, we should take seriously the possibility that two systems that solve it share deep explanatory similarities … difficult tasks are more constraining tasks, and success at difficult tasks justifies mechanistic/causal interpretations of our successful model". Thus, they suggest that for difficult enough tasks the realm of possible solutions may be constrained enough that any two algorithms that can solve this task are likely to exhibit 'deep explanatory similarities'. Even without careful etiological considerations, scientists may suggest the 'right' model. This is an interesting suggestion. But it seems to me that it is motivated by empirical results of correlations between simulated and neuronal activity that are related to object classification tasks. As I argued in 4A, however, such results do not show that the same computation is taking place in those two systems. Moreover, some counterexamples come to mind. Chess-playing seems like a difficult enough task, yet it is believed that 'deep-blue' solves it in a different manner than people. Finally, the argument in this paper is exactly that the functions the brain and the model are optimized to perform are different, while the latter is optimized for the function, the former may only *perform* it, without being optimized for it specifically. Therefore, the computations performed are likely to differ between the brain and the model.

5. Some concluding remarks

This paper argued that scientists must be more careful when appealing to etiological considerations when using the performance-based methodology. This is because not every capacity can be reasonably considered to be the result of a distinct optimizing process and treating it as such is likely to miss important aspects of how it is performed. Two main issues are worth emphasizing. First, although neuronal data can certainly be used to guide scientific search for the computations the brain performs, it is not a deciding factor. For neuronal correlations and causal relations can be identified for a variety of competing hypotheses about computations. Second, the fact that a function increases or increased the fitness of an organism does not mean that this is the result of a distinct optimization process, as demonstrated for the case of object recognition. In general, to discover what the brain computes, scientists should be sensitive to the manner in which the computations became possible and be much more careful in assigning histories to cognitive capacities. Without such caution, discovering computations in the brain is not necessarily impossible, but vastly more difficult.

References

Ashida, Go, and Catherine E. Carr. 2011. “Sound Localization: Jeffress and Beyond.” *Curr Opin Neurobiol.* 21: 745–51.

Atkinson, Anthony P, and Michael Wheeler. 2004. “The Grain of Domains: The Evolutionary-Psychological Case Against Domain-General Cognition.” *Mind & Language* 19(2): 147–76. https://doi.org/10.1111/j.1468-0017.2004.00252.x.

Banino, Andrea et al. 2018. “Vector-Based Navigation Using Grid-like Representations in Artificial Agents.” *Nature* 557(7705): 429–33. https://doi.org/10.1038/s41586-018-0102-6.

Boorse, C. 2002. “A Rebuttal on Functions.” In *Functions: New Essays in the Philosophy of Psychology and Biology*, eds. A Ariew, Robert C. Cummins, and Mark Perlman. Oxford University Press.

Bowers, J. S. et al. 2023. “Deep Problems with Neural Network Models of Human Vision.” *Behavioral and Brain Sciences*. https://doi.org/10.31234/osf.io/5zf4s.

Campbell, Murray, A.Joseph Hoane, and Feng-hsiung Hsu. 2002. “Deep Blue.” *Artificial Intelligence* 134(1): 57–83. https://www.sciencedirect.com/science/article/pii/S0004370201001291.

Cao, Rosa, and Daniel L.K. Yamins. 2022a. “Explanatory Models in Neuroscience: Part 1--Taking Mechanistic Abstraction Seriously.” *arXiv*.

———. 2022b. “Explanatory Models in Neuroscience: Part 2 - Constraint-Based Intelligibility.” *arXiv*.

Craver, Carl F. 2013. “Functions and Mechanisms: A Perspectivalist View.” In *Functions: Selection and Mechanisms.*, ed. Huneman P. Springer.

Cross, Logan, Jeff Cockburn, Yisong Yue, and John P O’Doherty. 2021. “Using Deep Reinforcement Learning to Reveal How the Brain Encodes Abstract State-Space

Representations in High-Dimensional Environments." *Neuron* 109(4): 724–38. https://www.sciencedirect.com/science/article/pii/S0896627320308990.

Cueva, Christopher J., and Xue-Xin Wei. 2018. "Emergence of Grid-like Representations by Training Recurrent Neural Networks to Perform Spatial Localization." In *International Conference on Learning Representations*,.

Cummins, Robert. 1975. "Functional Analysis." *The Journal of Philosophy* 72: 741–65.

Elber-Dorozko, Lotem, and Yonatan Loewenstein. 2018. "Striatal Action-Value Neurons Reconsidered." *eLife* 7: e34248.

———. 2023. "Do Retinal Neurons Also Represent Somatosensory Inputs? On Why Neuronal Responses Are Not Sufficient to Determine What Neurons Do." *Cognitive Science* 47(4): e13265. https://doi.org/10.1111/cogs.13265.

Fox, Lior, Ohad Dan, Lotem Elber-Dorozko, and Yonatan Loewenstein. 2020. "Exploration: From Machines to Humans." *Current Opinion in Behavioral Sciences* 35: 104–11. https://www.sciencedirect.com/science/article/pii/S2352154620301236.

Gibson, J.J. 1979. *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Goldstein, Ariel et al. 2022. "Shared Computational Principles for Language Processing in Humans and Deep Language Models." *Nature Neuroscience* 25(3): 369–80. https://doi.org/10.1038/s41593-022-01026-4.

Gould, S J, and R C Lewontin. 1979. "The Spandrels of San Marco and the Panglossian Paradigm: A Critique of the Adaptationist Programme." *Proceedings of the Royal Society of London. Series B, Biological Sciences* 205(1161): 581–98. http://www.jstor.org/stable/77447.

Hasson, Uri, Samuel A Nastase, and Ariel Goldstein. 2020. “Direct Fit to Nature: An Evolutionary Perspective on Biological and Artificial Neural Networks.” *Neuron* 105(3): 416–34. https://www.sciencedirect.com/science/article/pii/S089662731931044X.

Jonas, Eric, and Konrad Paul Kording. 2017. “Could a Neuroscientist Understand a Microprocessor?” *PLoS Comput Biol* 13: e1005268.

Kell, Alexander J E et al. 2018. “A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy.” *Neuron* 98(3): 630-644.e16.

Krakauer, John W et al. 2017. “Neuroscience Needs Behavior: Correcting a Reductionist Bias.” *Neuron* 93(3): 480–90.

Machery, Edouard. 2007a. “Discovery and Confirmation in Evolutionary Psychology.” In *The Oxford Handbook of Philosophy of Psychology*, Oxford University Press.

———. 2007b. “Massive Modularity and Brain Evolution.” *Philosophy of Science* 74(5): 825–38. http://www.jstor.org/stable/10.1086/525624.

Marom, Shimon et al. 2009. “On the Precarious Path of Reverse Neuro-Engineering.” *Frontiers in Computational Neuroscience* 3.

Marr, David. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. MIT Press.

McCaffrey, Joseph B. 2023. “Evolving Concepts of Functional Localization.” *Philosophy Compass* 18(5): e12914. https://doi.org/10.1111/phc3.12914.

Millikan, Ruth Garrett. 1989. “In Defense of Proper Functions.” *Philosophy of Science* 56(2): 288–302. http://www.jstor.org/stable/187875.

Mineault, Patrick, Shahab Bakhtiari, Blake Richards, and Christopher Pack. 2021.

“Your Head Is There to Move You around: Goal-Driven Models of the Primate Dorsal Pathway.” In *NeurIPS*,.

Neander, Karen. 1991. “Functions as Selected Effects: The Conceptual Analyst’s Defense.” *Philosophy of Science* 58(2): 168–84. https://doi.org/10.1086/289610.

Novick, Rose. 2023. *STRUCTURE AND FUNCTION*. Cambridge University Press.

Piccinini, Gualtiero. 2015. *Physical Computation: A Mechanistic Account*. Oxford University Press.

Proffitt, Dennis R. 2006. “Embodied Perception and the Economy of Action.” *Perspectives on Psychological Science* 1(2): 110–22. https://doi.org/10.1111/j.1745-6916.2006.00008.x.

Quartz, Steve R. 2002. “Toward a Developmental Evolutionary Psychology: Genes, Development, and the Evolution of the Human Cognitive Architecture.” In *Evolutionary Psychology: Alternative Approaches*, eds. Steven J. Scherand and Frederick Rauscher. Dordrecht: Kluwer, 185–210.

Schrimpf, Martin et al. 2020. “Integrative Benchmarking to Advance Neurally Mechanistic Models of Human Intelligence.” *Neuron* 108(3): 413–23. https://www.sciencedirect.com/science/article/pii/S089662732030605X.

———. 2021. “The Neural Architecture of Language: Integrative Modeling Converges on Predictive Processing.” *Proceedings of the National Academy of Sciences* 118(45): e2105646118. https://doi.org/10.1073/pnas.2105646118.

Shagrir, Oron. 2022. *The Nature of Physical Computation*. Oxford University Press.

Shteingart, Hanan, Tal Neiman, and Yonatan Loewenstein. 2013. “The Role of First Impression in Operant Learning.” *Journal of Experimental Psychology: General* 142: 476–88.

Sprevak, Mark. 2018. “Triviality Arguments about Computational Implementation.”

In *Routledge Handbook of the Computational Mind*, eds. Mark Sprevak and Matteo Colombo. London: Routledge, 175–91.

Turner-Evans, Daniel B et al. 2020. “The Neuroanatomical Ultrastructure and Function of a Biological Ring Attractor.” *Neuron* 108(1): 145-163.e10. https://www.sciencedirect.com/science/article/pii/S0896627320306139.

Wouters, Arno. 2005. “The Function Debate In Philosophy.” *Acta Biotheoretica* 53: 123–51.

Yamins, Daniel L.K., and James J. DiCarlo. 2016. “Using Goal-Driven Deep Learning Models to Understand Sensory Cortex.” *Nature Neuroscience* 19(3): 356–65.

Yamins, Daniel L K et al. 2014. “Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex.” *Proceedings of the National Academy of Sciences* 111(23): 8619 LP – 8624. http://www.pnas.org/content/111/23/8619.abstract.

Zhuang, Chengxu et al. 2021. “Unsupervised Neural Network Models of the Ventral Visual Stream.” *Proceedings of the National Academy of Sciences* 118(3): e2014196118. https://doi.org/10.1073/pnas.2014196118.