Can neuroscientists ask the wrong questions? On why etiological considerations are essential when modeling cognition

Lotem Elber-Dorozko

Abstract: It is common in machine-learning research today for scientists to design and train models to perform cognitive capacities, such as object classification, reinforcement learning, navigation 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 a worthwhile scientific finding that the workings of these models correlate with neuronal activity, as several prominent papers reported. This is a promising method to understanding cognition. However, I argue that, to the extent that this method's aim is to explain how cognitive capacities are performed in the brain, it is expected to succeed only when the modeled capacities are such that the brain has a sub-system dedicated to their performance. This is likely to occur when the modeled capacities are the result of a distinct adaptive or developmental process.

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, 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; Schrimpf et al. 2021), navigation[1] (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 closely resembles Marr's (1982) levels of analysis: it begins by describing the performed computation, then it identifies an algorithm which can perform this computation, and finally it searches for the algorithm's neuronal correlates. This approach emphasizes the usefulness of top-down constraints in modeling neuronal activity – the model must be able to perform the cognitive function in which the brain area is involved. At least in the step of constructing the algorithm for the cognitive capacity, this approach also minimizes the

[1] 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.

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. This leads some to suggest that the constraint that the algorithm must perform a specific cognitive capacity may be sufficient to create similarities in the algorithm utilized by the machine-learning algorithm and in cognitive processing (Cao and Yamins 2022a; Yamins and DiCarlo 2016).

Here, I argue that, to the extent that this methodology aims to explain how cognitive capacities are performed, it must take into account another aspect of cognition, over and above neuronal activity and behavior; we must also consider whether the capacity it aims to model is a capacity for which there is a dedicated 'sub-system' in the brain. For if this is not the case, it is very unlikely that scientists will be able to model how said capacity is performed in the brain. If the modeled capacity is not performed by a dedicated sub-system in the brain, machine learning models will overlook important constraints on the way the capacity is performed. Here, 'Sub-system' is simply taken to mean that there are specific properties and states of the brain that are dedicated solely to the performance of a specific capacity. Such 'sub-systems' need not be anatomically distinct. Despite the difficulty in identifying specific capacities in biology (Wouters 2005), some well-known examples for sub-systems in the body include (simplistically described) various organs in the body, such as the heart – dedicated to pumping blood, the ribosome – dedicated to building proteins, or the immune system – dedicated to preventing or limiting infection.

This paper further argues that to identify sub-systems in the brain scientists should pay careful consideration to the etiology of cognitive capacities. This, because

capacities that are not the result of specific adaptations (or developmental/learning processes if one considers the brain to be plastic enough) are very unlikely to have sub-systems dedicated specifically to them.

Finally, 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 that a specific computation is performed in the brain. 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 completely eliminated, even with much empirical data about neuronal activity. Therefore, putting too much weight on correlational data while ignoring etiological considerations, may lead to erroneous attribution of computation to the brain.

The emphasis this paper suggests on evolutionary considerations is not novel. It has been repeatedly suggested by neuroscientists and philosophers that cognitive scientists should pay mind to evolutionary processes. In a special issue dedicated to neuroscience and evolution, Cisek and Hayden (2022) write: 'we think that the consideration of evolutionary history ought to take its place alongside other intellectual tools used to understand the brain'.[2] Moreover, the performance-based methodology this paper addresses is regularly taken to be the champion of evolutionary considerations. For, often, this methodology results in neuronal activity

[2] It should be noted, that Cisek focuses more on phylogenetic analyses, while this paper advocates for adaptationist considerations.

being described as activity that has been adapted for a specific purpose – the one that the algorithm was trained on. This description often supplants previous descriptions of neuronal activity as responding to well-defined concepts such as oriented lines or faces. Indeed, Cao and Yamins (2022b) write: "we want to find tasks that are good proxies for evolutionary goals that brains were actually selected for achieving."

This paper aims to add onto this literature in several novel ways. First, it demonstrates concretely how etiological (used here to simply mean evolutionary or developmental) considerations are relevant when using a specific method to identify neuronal computation. Second, it argues for the importance of etiological considerations, not only as a heuristic tool to identify possible models for computation in the brain but as playing a role in the assessment of these models together with neuronal and behavioral data. This paper argues that decisions about computations in the brain are likely to depend on an interplay of etiological considerations, together with empirical data about brain activity, brain structure and behavior. This interplay has generally been overlooked, with scientists commonly appealing to experimental data alone to justify their conclusions. Finally, this paper argues that we must adopt a more nuanced view of etiology. The fact that having a capacity increases an organism's fitness does not mean this is a capacity that the organism has adapted to have a specific sub-system for. Ignoring such distinctions may lead to erroneous attribution of computations 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 capacities for which there are dedicated sub-system. It will make this argument by demonstrating three ways in which failure to do so can lead us astray. Then, this section will discuss the difficulties this argument raises for current

scientific practice and will suggest a way forward. Finally, section 4 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) train a model that can perform an object classification task at near human performance; The model can classify objects from 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) posit about neuronal processing (the function performed by the neurons is some linear operation on neuronal activity in the previous layer, 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; the model was only trained to perform the task as best as possible, and information about neuronal activity was not used during training.

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 computational model was able to predict activity in the inferior temporal cortex (IT) - 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. 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).

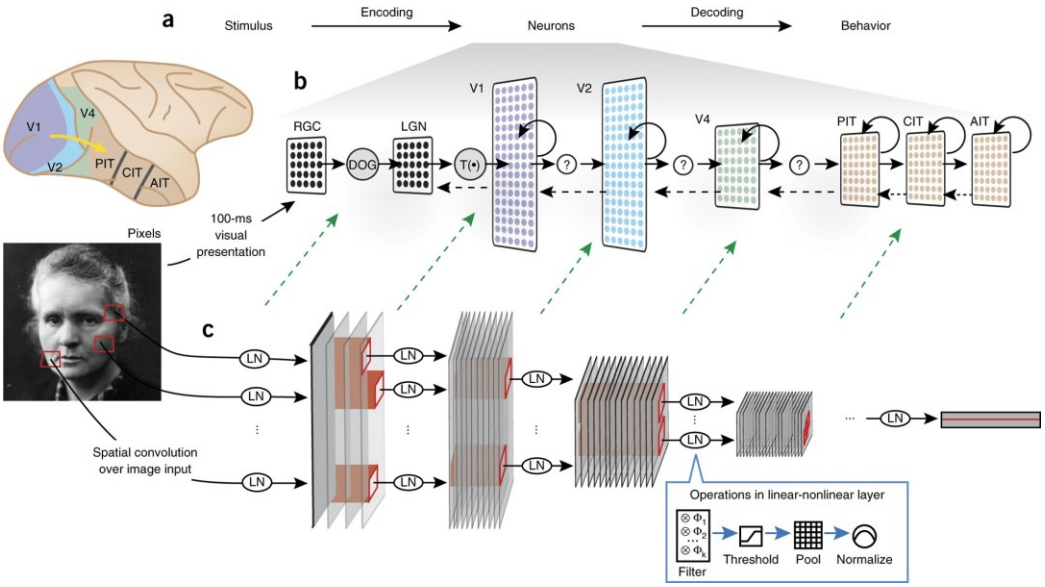

Fig. 1, from (Yamins and DiCarlo 2016), a performance-based model for object classification as a model of computation in the brain. Each layer in the model is mapped to an area in the brain, with corresponding processing stages.

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 the impressive results 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 this methodology is likely to yield explanation and understanding of cognition only when it attempts to explain capacities for which there is a dedicated sub-system. I further argue that one important way to identify these capacities is to consider their etiology. Without such etiological considerations, although simulated activity may show some mapping to neuronal activity, the computational models are likely to be different in important ways from the ones

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

employed by the brain, because they will miss important constraints on how the capacity is performed. I elaborate in the next section.

3. How etiological considerations matter

An important advantage of the performance-based methodology is that its models are able to perform cognitive capacities. This makes it much more likely that they capture the essence of how brains perform those capacities, compared with models that cannot perform said capacities. Nonetheless, another important aspect to consider is the plausibility that the brain performs the capacity in the same manner the model does. The next three sub-sections describe three different ways in which the process underlying a capacity is evidently not a dedicated sub-system, leading to models that substantially diverge from the true processes underlying the modeled capacity.

A. Modeling side effects, rather than adapted functions

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 capacities the brain has adapted to have 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 have a sub-system dedicated to their performance and therefore 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, and 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. 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 models they will come up with 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.[4] Similar points can be made with regard to driving, baking a cake and synchronized swimming. While some of these capacities are useful today, such as driving, it is clear that this capacity exists not as the result of a dedicated sub-system, but as a side-effect of our perceptual and motor abilities more generally. Thus, attempting to create models that can perform these

[4] One may suggest that chess experts develop specific mechanisms that are not constrained by other tasks, to support chess playing. This is not impossible, but does not fit with what is known about neuronal processing or about how acquisition of skills in chess affect the brain (Mayeli, Rahmani, and Aarabi 2018), and at any rate at least novice chess players are very unlikely to have such a module.

capacities alone is unlikely to yield similar processing to how they are performed in the brain. Modeling capacities that are evolutionary side-effects is unlikely to yield models similar to computation in the brain.

B. Modeling overly specific capacities

Not every capacity that is considered a function according to the selected-effects view will have a dedicated sub-system. 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."

Machery's paragraph nicely captures the notion of Darwinian modules. One may wonder if the examples given by Machery should be considered side-effects and belong in sub-section 3a, since they are too recent to be the result of an adaptive process. Nonetheless, similar arguments can be made for capacities that are not side-effects, but simply partial descriptions of capacities with dedicated sub-systems, and for this reason do not have a dedicated sub-system.

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' and have a dedicated sub-system. 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 must perform a much more complex capacity, not only to distinguish zebras from tigers, but also to distinguish zebras from cats, chess boards, and cross walks. Thus, because the scientist chose an overly narrow description of a capacity, one that the brain has not adapted to have a dedicated sub-system for, it is very unlikely that 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. The upshot is that attempting to use performance-based modeling to explain any capacity that is only a partial description of a capacity with a dedicated sub-system is expected to miss important constraints on how the capacity is performed and is unlikely to result with explanatory models.

When discussing evolutionary psychology, one should mention the debates surrounding this practice. Evolutionary biology in general has been criticized as prone to telling 'just-so' stories, making claims that sound compelling about how certain features serve certain functions without evidence or justification (Gould and Lewontin

1979). There is also debate on the extent to which cognition can be characterized as evolutionary modular – the extent to which selection pressures can be separated for different cognitive capacities (Machery 2007b; Quartz 2002). A third related issue is the grain problem – how should scientists know what is the right level of specificity of which to define a cognitive capacity? Atkinson and Wheeler (2004) nicely point out that there is no area of research that can fully answer this question, as both phenotypes and problems of adaptation can be treated at different levels of granularity. Finally, even if scientists correctly identify specific selection pressures, there is no guaranty that adaptation will lead to a capacity that is well-designed to perform the specific task. I will go a bit more in depth into these issues at section 3E. For now, it will suffice to point out that the approach suggested here focuses on the implausibility of certain claims given what we know about evolution, and does not aim to support tenuous hypotheses. Therefore, it is not in much danger to tell 'just-so' stories. Moreover, much work in evolutionary psychology centers around very high level cognitive capacities – mate choice, recognition of cheaters, etc. (Machery 2007a). These are areas for which it will be very difficult to identify other converging evidence for a distinct sub-system for the investigated capacity. Performance-based methodology focuses on more basic capacities – object classification, decision-making, navigation. For such capacities there is more room for neuronal and etiological consideration to constrain each other. Finally, despite all the challenges presented for evolutionary modularity, for certain capacities it is quite clear that they are not capacities for which there are dedicated sub-systems, such as driving or differentiating zebras and tigers. This knowledge can be important for scientific practice.

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 entail 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 to 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 will have a dedicated sub-system 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, or they are both. This is the rubric that describes many capacities modeled in scientific practice.

D. Current scientific practice

While current scientific 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 for which there is a dedicated sub-system.

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 the brain does not have a dedicated sub-system. 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 to perform this object classification 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 there is a dedicated sub-system to classify objects from images. Perhaps it is more reasonable to say that perception adapted 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. 2022) 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 relatively immediate rewards. Therefore, one could call the ability to play video games a side-effect of the human capacity for decision-making, and 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 have a dedicated sub-system in the human brain that deals with complex, changing, and continuous environments.

The claim that current scientific practice does not model capacities with dedicated sub-systems is not meant as a criticism of 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 include 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. To be a plausible model for a cognitive capacity the modeled capacity should be one which is likely to have a dedicated sub-system.

E. Identifying sub-systems

So far, this paper has described several cases where, intuitively, the modeled capacity is not a capacity for which there is a dedicated sub-system and therefore do not expect the model to correspond to computation in the brain. The question that obviously arises is how one can identify these sub-systems.

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. Although this issue stands, it is an empirical fact that we can divide many biological systems into subsystems in a way that aids in 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.

Moreover, there are parts in the brain that, although not all is understood about how they work, are already considered dedicated sus-systems for specific capacities. 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 sub-system that implements the 'Jeffress model' to compute interaural time difference, the time delay of sound between the two ears (Ashida and Carr 2011). This computation is the basis for sound localization for certain frequencies. As another example, there is strong evidence that an area in the central complex of the fly brain implements the computational model

known as the ‘ring attractor’ to represent head direction (Turner-Evans et al. 2020). The upshot of all these examples is simply that it is probable that we should be able to identify other sub-systems in the brain and should not give up on this mission, even when we know in advance that there is no such thing as a completely independent, distinct evolutionary pressure. It may turn out for some general capacities, such as decision-making or language processing, that they cannot be divided into simpler sub-capacities with dedicated sub-systems. Then, the consequence is that the modeled capacities should not be divided as well. Modeling such complex capacities would be an extremely challenging task for scientists. Nonetheless, it may be what should be done to explain these cognitive capacities.

How will such sub-systems be identified? I suggest that to identify computation in the brain we should appeal to an interplay of evidence as well as common-sense assumptions from different domains. Attempting to model capacities for which dedicated sub-systems are more plausible 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 come to perform using a dedicated sub-system. 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 sub-systems. However, etiological considerations never disappear entirely, even a 99% prediction of neuronal activity would not convince us that the brain is calculating the location of Mars relative to Neptune. Thus, there is reciprocity between considerations of 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, but etiological considerations aid in constraining the models that scientists consider. Sciences would be very lucky to come up with the right model for how the brain performs a capacity, when their model is optimized specifically for this capacity, while the brain has adapted to have different capacities and does not have a dedicated sub-system for the modeled capacity. The closer the capacity considered by scientists and the capacity the brain has a dedicated sub-system for, the more similar we can expect the scientific model and neuronal computation to be.

In the next section I present some objections to the claim that without etiological considerations scientists may be supporting the wrong models.

4. Some objections

A. Neuronal correlates can fully support a specific computation

One evident objection to the claim that computational modeling requires etiological considerations, is to note that this argument completely ignores the neuronal data. 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

[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."

completely different computation[6] (Elber-Dorozko and Loewenstein 2018; Jonas and 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', leading many scientists to believe that the computation of this variable is essential to decision-making. 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 discovered that standard analyses performed on this model still erroneously identified significant representation of 'action-value'.

These results demonstrate that correlation does not imply 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 enough data will result in a significant correlation. Thus, identification of a correlation between neuronal activity and some variable is

[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 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 constrained methods employed in neuroscience.

not an indication that this variable is computed, but only that neuronal activity is not completely orthogonal to this variable. Given that any computational variable that performs some capacity is likely to correlate with properties of the inputs and the outputs of the capacity to some magnitude, it would seem that there are many possible computational models that correlate with neuronal activity without being identical to neuronal computation.

Even though scientific results of correlation with neuronal activity cannot imply a specific computational model, still much can be learned from them. First, as long as 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. 2022) 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 is one that has a dedicated sub-system. But, if it is implausible that the modeled capacity has a dedicated sub-system, 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. The determinants of computation are not etiological

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 history of the system.

One could, of course, deny that etiological considerations are relevant when considering what the brain, and other systems, compute. There are several popular views of computation that do just that. 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. Most notably, it is pointed out that in biological systems the etiology is often unknown, and this does not stop scientists from assigning functions to systems. Moreover, it does not seem 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 the latter criticism, I note that this debate often centers 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 practically never happen, and first mutations tend to be small and to build upon previous states. Moreover, attempting to analyze a swamp-person from a scientific perspective seems a nearly hopeless endeavor. For if scientists will try to explain chess playing the same way they explain sound localization (Ashida and Carr 2011), as a unique module, they will have a very hard time making any progress. 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 relevant and useful for understanding how they perform the computations – organisms have evolutionary histories and computers are designed.

Another challenge to the claim that 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 determining computation, and nonetheless scientists move forward with assigning functions and computations. One answer to this is that while it is challenging to know the exact etiology of capacities, some general properties are quite easily considered, as can be seen in the work of evolutionary psychologists and implicitly in the choices of neuroscientists. To illustrate, we know that brains have not adapted for chess-playing or for stock-trading, or that they adapted for a specific mechanism for telling zebras and tigers apart. Neuroscientists certainly use these intuitions when choosing capacities to model. Considering various

organisms can also be telling about the etiology of capacities. Moreover, as described in section 3, models that are built to perform etiologically relevant capacities are also more likely to be similar to behavior and to neuronal activity. Therefore, while etiological considerations are not perfect, they are an important part of building more realistic models and being more explicit and clear about them can help scientists in explaining cognition.

Finally, I present a challenge to the 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 what a system computes 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 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, if neuroscientists wish to claim that their models aim to capture a specific capacity which was created with independent etiological pressures, 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 to the question of whether it is reasonable that this quick classification in the forward pass is the result of specific evolutionary pressures that yielded a sub-system.

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 with specific, independent, etiologies.

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'. Therefore, even without 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 imply 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. It still may be that for certain tasks the possible solutions are constrained enough, but this seems like an open, empirical question.

5. Some concluding remarks

This paper argued that scientists must take etiological considerations into account when using a performance-based methodology to model neuronal computation. This is because the history of how a computation came to be determines whether it will has a dedicated sub-system. 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. Instead, more weight should be given to etiological considerations. Second, the fact that a function increases or increased the fitness of an organism does not mean that

this is the right description of the function for which it has a dedicated sub-system, as demonstrated for the cases of object recognition. In general, to discover what the brain computes, scientists should be sensitive to the manner in which the computations became possible. Without such sensitivity, discovering computations in the brain is not 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. 2022. *Deep Problems with Neural Network Models of Human Vision*. 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*.

Cisek, Paul, and Benjamin Y Hayden. 2022. "Neuroscience Needs Evolution." *Philosophical Transactions of the Royal Society B: Biological Sciences* 377(1844): 20200518. https://doi.org/10.1098/rstb.2020.0518.

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.

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.* 205: 581–98.

Jonas, Eric, and Konrad Paul Kording. 2017. "Could a Neuroscientist Understand a

Microprocessor?" *PLoS Comput Biol* 13: e1005268.

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.

Mayeli, Mahsa, Farzaneh Rahmani, and Mohammad Hadi Aarabi. 2018. "Comprehensive Investigation of White Matter Tracts in Professional Chess Players and Relation to Expertise: Region of Interest and DMRI Connectometry ." *Frontiers in Neuroscience* 12. https://www.frontiersin.org/article/10.3389/fnins.2018.00288.

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.

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.