# Towards a common theory of explanation for artificial and biological intelligence

**Jessica A F Thompson (j.thompson@umontreal.ca)**
International Laboratory for Brain, Music and Sound (BRAMS)
Quebec Artificial Intelligence Institute (Mila)
University of Montreal
Montreal, Quebec, Canada

## Abstract

**Much of the confusion that occurs when working at the intersection of cognitive science, artificial intelligence, and neuroscience stems from disagreement about what it means to explain intelligence. I claim that to integrate these fields, we must reconcile their different theories of explanation. I briefly review theories of scientific explanation in neuroscience and recontextualize the stated views of several prominent cognitive computational neuroscientists in terms of the theories of explanation they espouse. Finally, I describe some of the challenges of forging a new theory of explanation that would apply equally to artificial and biological intelligence. As a first step towards an integration of research on biological and artificial intelligence, my goal in writing this paper is to equip scientists of intelligence to interrogate and justify the theories of explanation that underlie their definitions of scientific progress.**[1]

**Keywords:** philosophy of science; philosophy of mind; explanation; neuroscience; artificial intelligence; deep learning

## Integration of AI and neuroscience

Much lip service is paid to the integration of deep learning and neuroscience, with the goal of creating a feedback loop—deep learning for neuroscience and neuroscience for deep learning. However, this has proven difficult in practice. The confusion when trying to work at this intersection comes in large part from lack of agreement about what progress towards a common goal would look like. This topic came up at the inaugural Cognitive Computational Neuroscience (CCN) conference last year, which was assembled to unify the "disconnected communities of cognitive science, artificial intelligence, and neuroscience" towards the common goal of "understanding the computational principles that underlie complex behavior" (Naselaris et al., 2018). Jim DiCarlo, chairing a panel discussion, asked, "when people say they want to work together, usually there is some idea of a shared goal...some idea of what success would even look like...are we even after the same thing?" This question received a number of very different answers from the panel, demonstrating the challenge of even agreeing on a common goal. Panelist Yann LeCun stated the common goal to "explain intelligence" but this doesn't answer the question because we disagree about what it means to explain intelligence. LeCun wants to replicate animal intelligence in artificial systems. On the other hand, for neuroscientist Jackie Gottlieb, "success means characterizing a system at a particular level of abstraction ... in a way that is reproducible and solid" (Kay, 2017). Cognitive scientist Josh Tenenbaum, stressed the importance of distinguishing between goals on different time scales and suggested that all the CCN attendees probably share some long term vision of success, even if they disagree about what to do to work towards that goal in the short term. An integration of cognitive science, artificial intelligence and neuroscience will not be possible until we are able to motivate our research by reference to a shared definition of what it means to make progress towards the goal of explaining intelligence.

The same debate is happening in machine learning right now. The quest for interpretable AI is ultimately asking, What explanations of AI systems will we accept? Are some systems more explainable than others? For example, are systems that are designed specifically to expose 'disentangled' representations more interpretable? Several events were dedicated to related topics at the Neural Information Processing Systems conference in 2017 (e.g., Interpretable ML Symposium, Learning Disentangled Representations: from Perception to Control). It is no coincidence that machine learning and neuroscience are both having these conversations now. Rather, it is precisely because artificial systems are looking more like biological ones and our models of biological intelligence are looking increasingly like AI that we are forced to question our standard conceptions of what makes a good explanation.

These questions are ultimately in the realm of epistemology and philosophy of science, yet philosophical theories are not often (explicitly) invoked in the discussion. As scientists, instead of reinventing the wheel, we would do well to look to our philosopher colleagues to help us wade through these difficult but crucially important questions about what constitutes an explanation. At the very least, our discussion would be simplified if we borrowed from the established language of philosophy of science. I claim that what is needed is actually to create a new theory of explanation that applies equally to biological intelligence and artificial intelligence. The nature of research on AI (the methods we use to study AI, the nature of explanations that we accept) is very different from the way we traditionally conceptualize the study of biological systems. At present, this constitutes a challenge to the CCN goal, but in the long term, I see this as an opportunity to define a new science of intelligence that includes both artificial and biological intelligence. My central claim is that to achieve an integration of cognitive science, AI and neuroscience, we must reconcile their different theories of explanation. Background in the philosophy of

[1] An earlier version of this paper was presented at the 2018 Cognitive Computational Neuroscience Meeting (CCN2018).

scientific explanation of neural and computational systems will better equip us to express our views on questions like, How ought intelligence be explained?, and ultimately to be able to design experiments that we can rigorously justify.

## Theories of Explanation

To begin, we assume that a primary goal of science is to provide explanations of phenomena (be they natural, social or technological). The role of a theory of scientific explanation is to characterize the structure of explanations in science. An account of scientific explanation must distinguish between explanations that are scientific and those that are not. It must also distinguish between explanations and non-explanations. Sometimes this second contrast is presented as the difference between explanation and description. For example, a set of claims about the appearance of a particular species may be true, accurate and supported by evidence without being explanatory in any way. They are *merely* descriptive (Woodward, 2017).

Scientists ought be able to look at a set of claims or a model and judge whether it is descriptive or explanatory (or neither) according to one's favorite theory of scientific explanation. This is not to say that description is somehow inherently less valuable than explanation. We must know that a phenomenon exists before we can ever hope to explain it. In calling for better literacy of theories of explanation, my wish is not to reduce the amount of descriptive science, but rather to reduce the misrepresentation of scientific activities. If we already think that we are explaining, then we will not spend time figuring out how to explain. Instead, if we acknowledge that what we are currently doing is descriptive, we can better see our role in a larger scientific enterprise, i.e. the role of a particular series of experiments may be to describe a specific set of phenomena that may later be explained by other experiments. When our perceived explanatory power is inflated, we're closed off from seeing how we might work together to ultimately better explain in the long term.

## What do contemporary scientists say?

In this section I will discuss how some computational neuroscientists have recently answered the question, What makes a good model (or theory) of the brain? In their answers, we will find their philosophical commitments.

In his talk *Playing Newton: Automatic Construction of Phenomenological, Data-Driven Theories and Models*, Ilya Nemenman refutes the claim that a good theory of brain must be a large, multi-scale computational model with a quote from Rosenblueth Wiener, "Theories must lose details and must be developed to explain limited sets of phenomena. Otherwise, the best material mode of a cat is another, or preferably the same, cat" (Rosenblueth & Wiener, 1945). In other words, "Don't model bulldozers with quarks" (Goldenfeld & Kadanoff, 1999). A good theory is one that accurately explains a limited set of phenomena and throws away everything is wasn't designed for (Nemenman, 2018). But what does it mean to *explain* to Nemenman? He suggests that we can ignore philosophical answers to this question because philosophy has failed to define science. In his view, we don't need philosophy anyway because we have Bayesian statistics. Bayesian statistics can already tell us what is falsifiable and what is falsified. If a theory does not explain the observed data, then the theory is falsified. If a theory can explain any dataset, then it is unfalsifiable. What makes a good model of the brain is determined by Bayesian model selection. The goodness of a theory is related to generalization and prediction, with little regard to whether it is true. Falsifiability is then real-valued and empirical. If it explains the data, we don't care if it is correct or not. According to Nemenmen, good models:

- are phenomenological,
- predict your data and generalize to new data and experiments, and
- only explain the specific question they were designed to answer with as few parameters as possible.

This view reflects the well known heuristic of Occam's Razor: the best solution is the one that explains the data best with the simplest model.

Neumenman advocates for phenomenological models, which have been characterized by philosophers of science as black box models because they merely capture input-output relationship for the phenomena to be explained without positing intervening variables or mechanisms. In computational neuroscience, phenomenological models are called descriptive models because they "summarize data compactly" without addressing "the question of how nervous systems operate on the basis of known anatomy, physiology, and circuitry" (Dayan & Abbott, 2005). Nemenman claims that the important test is whether the model generalizes to unseen data and new situations (Nemenman, 2018).

Consider the Balmer formula as a counter example, which was constructed to describe the four visible lines in the emission spectrum of hydrogen. Not only does this model accurately capture the target phenomena, but it also successfully predicted the existence of previously unobserved spectral lines outside of the visual range. Despite its success, none of the model elements have any physical interpretation. Balmer arrived at this formula via trial and error curve fitting to find the best mathematical fit to the four visible spectral lines of hydrogen. It is generally agreed that the Balmer formula is not explanatory because it does not explain why the emission spectrum for hydrogen shows the pattern that it does (Kaplan, 2011). Phenomenological models can certainly be very useful and are an important part of any scientific enterprise, but to claim that they are explanatory is a difficult claim to defend.

In *Principles for models of neural information processing* (?), Kendrick Kay defines cognitive neuroscience to be the quest for explanations of the mind in terms of the brain. Yet, in his definition of explanation, "models posit that specific variables relate to neural activity. As such, models provide explanations of measurements of the brain." Is the phenomena

to be explained the behaviour (or perception) of the animal or the neural activity itself? The implication seems to be that if we can explain the neural activity, this will imply an explanation for the behaviour or perception. Consider the example given in the paper, Why is a neuron highly active during a clip of rock music? We can compare different models (the neuron responds to sound intensity versus the neuron responds to guitar tones) and evaluate which model is more accurate. For example, we can vary the intensity of a variety of sounds and find that the activity of this neuron scales with the sound intensity, while not being selective to any other tested features. Have we then explained the activity of this neuron and at the same time, explained perception of sound intensity? Some philosophers claim that explanations of human brains proceed via this type of *functional analysis*,"according to the explanatory strategy of functional analysis, the overall behavioral capacities [are] explained by breaking down or decomposing the capacities into a number of 'simpler' subcapacities and their functional organization" (Kaplan, 2011). This theory of explanation goes hand in hand with the information processing theory of mind. When you assume that the brain is best understood as an information processing machine, then once you've discovered the signals that it sends around and the operations that it performs, then you should be able to explain the resulting behaviour. Under a different theory of mind (e.g. embodied dynamicism), or a different theory of explanation (e.g. mechanistic), functional analysis is not explanatory. However, many mechanists would say that functional analysis is a necessary step en route to a mechanistic explanation.

Jonas Kubilius, also asks the question, What does it mean to understand? in his commentary "Predict, then simplify". To Kubilius, predictive power is the first and foremost attribute on which to assess a model. He and the rest of Jim DiCarlo's lab appear to be staunch predictivists, taking an engineering approach to neuroscience. This has led their group to find that deep neural network models trained originally for visual object recognition, better predict activity in much of visual cortex better than previous scientist-designed models (Yamins & DiCarlo, 2016). Predictivism refers to the view that phenomenological models are explanatory by virtue of their descriptive and predictive power. The problems facing predictivism are well known in philosophy of explanation. "Simple examples readily expose how accurate prediction is insufficient for explanation, and so how the predictive import of a given model can and often does vary independently of its explanatory force. On can accurately predict a storm's occurrence from a dropping barometer, but this does not explain the occurrence of the storm. Rather, a common cause—a drop in atmospheric pressure—explains both the dropping barometer value and the approaching storm. Similarly, a p-model might be predictively adequate, and yet its variables might only represent factors that are mere behavioural correlates of some common or joint cause for the target phenomenon. Just as we reject the claim that the barometer drop explains the storm, we should also resist the claim that p-models of this kind provide explanations" (Kaplan, 2011, p.351). A defense of predictivism must ultimately address these types of counter examples.

On first read of Kay an Kubilius, one might think that their disagreement about DNNs being explanatory models of the brain comes from placing different values on the predictive power of a model. Kubilius says predictive power comes first and foremost. Where as Kay says that the model must first be understood before it can explain. Yet both rely largely on prediction of observational data to validate their models. The larger disagreement is about the form of a good model. Both Nemenman and Kay ascribe to the heuristic of Occam's Razor, that the simplest model that explains the data is best. They want the number of components to be small because ultimately their modelling efforts are about finding which components are important. Kubilius points out that although historically, science's greatest successes have been the result of describing complex systems with relatively few parameters, there is no reason to assume that this will be the case for understanding the brain. Others have pointed out the absurdity of Occam's Razor, "how could a fixed bias toward simplicity indicate the possibly complex truth any better than a broken thermometer that always reads zero can indicate the temperature? You dont have to be a card-carrying skeptic to wonder what the tacit connection between simplicity and truthfinding could possibly be" (Kelly, 2007). This adherence to Occam's razor is part of what separates Neumenman and Kubilius. Simpler things seem to be easier for us to understand, so the tendency of science to look for simplicity when it is available makes sense. Kubilius suggests that we may look for simple processes that generate complex systems instead. For example, training deep neural networks depends on a number of principles (e.g. optimization of cost functions via gradient-based methods, compositionality, distributed representations) that is likely smaller than the number of parameters in the model.

## Towards a new theory of explanation for a new science of intelligence

There is a growing body of work trying to understand deep learning using empirical methods that look almost neuroscientific: ablation analyses, receptive field analysis, psychophysics. Nikolas Kriegeskorte calls it "synthetic neurophysiolog" (Kriegeskorte, 2015), Jeff Clune calls it "artificial neuroscience" (Metz, 2018), and Maithra Raghu calls it "deep learning scienc" (personal communication, Dec 7 2017), presumably because it involves applying the scientific method to study artificial systems. What I find most exciting about this line of inquiry is that it makes many connections to deep learning theory. If we can relate deep learning theory to empirical analyses of deep learning systems, then might there also be the potential to relate our empirical analyses of biological intelligence to a similar mathematical understanding? On the other hand, Eric Jonas and Konrad Kording asked, Could a neuroscientist understand a microprocessor? in their work which applied common neuroscientific analyses to electrical

measurements of a microprocessor. They show that although these analyses were able to make replicable and reliable descriptions of patterns of activity, they did not ultimately reveal the known and accepted explanation of how the microprocessor works (Jonas & Kording, 2016). Swapping methodologies, approaches and philosophies between deep learning and neuroscience has the potential to demonstrate the strengths and limitations of our scientific activities and perspectives, helping us to select those that will be most useful towards our common goal of understanding intelligence.

If the nature of a good scientific explanation is dependent on the phenomenon to be explained, then, to the extent that an artificial system and a biological system demonstrate the same phenomenon, their explanations should share the same form. All contemporary theories of explanation in neuroscience focus on physical computation. The mechanistic account, which currently dominates philosophy of neuroscience, requires that explanations consist of physical components in causal relationships with one another. The explanations we have for deep learning are concerned with abstract computation. The same principles hold regardless of which type of GPU your model is trained on. Reconciliation of these views suggests a commitment to multiple realizability. However multiple realizability is often associated with functionalism and computational chauvinism. We want a version of multiple realizability for the 21st century aligned with a modern theory of mind.

We know from machine learning research that there are many equivalent solutions to the optimization problems posed in deep learning. Repeated optimizations of the same network lead to solutions that occupy distinct regions in function space. The loss surface contains many equally good local minima (Erhan, Courville, & Vincent, 2010). Thus, in deep learning, specifically characterizing the exact function learned by the network is not very informative. As such, function identification will not play a large role in a unified theory of explanation for biological and artificial intelligence.

## Conclusion

A new theory of scientific explanation for both artificial and biological intelligence ought:

1. reflect that learning is central to intelligence,
2. imply multiple realizability without computational chauvinism,
3. abandon the focus on physical computation, and
4. not be concerned with characterizing the specific function that is computed by a network.

If our ultimate goal is truly to explain intelligence, then we must eventually agree on a theory of explanation that accounts for successful explanations of both artificial and biological intelligence. Although the mechanistic framework is currently very popular in most of philosophy of science, it is difficult to apply to the explanation of cognitive phenomena. Hence, much of cognitive and computational neuroscience that claims to be explanatory relies on older problematic theories of explanation and outdated theories of mind. To forge a new path forward, we need to first acknowledge that we fundamentally do not know how to explain intelligence. Until we do, let us be explicit about our philosophical commitments and let us value our descriptions without mistaking them for explanations.

## References


Dayan, P., & Abbott, L. (2005). *Theoretical neuroscience: Computational and Methematical Modeling of Neural Systems*. MIT Press.

Erhan, D., Courville, A., & Vincent, P. (2010). Why Does Unsupervised Pre-training Help Deep Learning? *Journal of Machine Learning Research*, *11*, 625–660.

Goldenfeld, N., & Kadanoff, L. P. (1999). Simple lessons from complexity. *Science*, *284*(5411), 87–89.

Jonas, E., & Kording, K. P. (2016). Could a neuroscientist understand a microprocessor ? *bioRxiv*, 055624.

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

Kay, K. N. (2017). *Opening panel discussion (CCN 2017).* Retrieved from `https://www.youtube.com/watch?v=3YhJW4Fp1xQ`

Kelly, K. T. (2007). Simplicity , Truth , and the Unending Game of Science. In S. Bold, Benedikt Lowe, T. Rasch, & J. v. Benthem (Eds.), *Foundations of the formal sciences v: Infinite games.*

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modelling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

Metz, C. (2018, 3). *Google Researchers Are Learning How Machines Learn.* https://www.nytimes.com/2018/03/06/technology/google-artificial-intelligence.html.

Naselaris, T., Bassett, D. S., Fletcher, A. K., Kording, K., Kriegeskorte, N., Nienborg, H., ... Kay, K. N. (2018). Cognitive Computational Neuroscience: A New Conference for an Emerging Discipline. *Trends in Cognitive Sciences*, *22*(5), 365 - 367.

Nemenman, I. (2018). *Playing Newton: Automatic Construction of Phenomenological, Data-Driven Theories and Models.* https://simons.berkeley.edu/talks/ilya-nemenman-4-17-18.

Rosenblueth, A., & Wiener, N. (1945). The Role of Models in Science. *Philosophy of Science*, *12*(4), 316–321.

Woodward, J. (2017). Scientific Explanation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (https://pl ed.). Metaphysics Research Lab, Stanford University.

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.