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


Jessica A.F. Thompson

International Laboratory for Brain, Music and Sound Research
Montréal, Canada

Montreal Institute for Learning Algorithms (Mila)
Montreal, Canada

Centre for Research on Brain, Music and Language (CRBLM)
Montréal, Canada

j.thompson@umontreal.ca


## Introduction

A primary goal of science is to provide explanations of phenomena, be they natural, social, or otherwise (Many other goals: forecasting, diagnosis, technology development). Ideally, the choice of scientific actions should rely on some road map for how to make progress towards at least one scientific goal. When the goal is to explain, this road map requires some idea of what type of explanation we're looking for.

There is little agreement or formalization of what would be acceptable explanations of the phenomena of intelligence. Panelists at CCN2017 gave varied answers to the question "what is your definition of success?" The quest for interpretable machine learning is ultimately a similar question, what explanations of AI systems will we accept?

**Claim**: The integration of deep learning and neuroscience will require a common theory of explanation that applies to both artificial and biological intelligence
**Goal**: Equip scientists of intelligence to interrogate and justify the theories of explanation that underlie their definitions of scientific progress

### The role of a theory of scientific explanation

- Characterize the structure of explanations in science
- Distinguish between explanations that are scientific and those that are not
- Distinguish between explanations and non-explanations. Sometimes 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.
- Reveal criteria for assessing explanations

**Phenomenological models** (aka "descriptive models"): describe or "save" a phenomena, summarize data compactly

Descriptive and normative goals of this field of philosophy:
- Descriptive: characterize explanations in contemporary science
- Normative: clarify the distinction between good and bad explanations

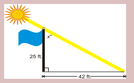
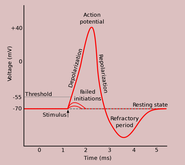
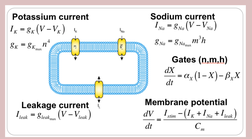
### Particular challenges to explaining intelligence

The form of a "good" explanation of a given phenomenon depends on the nature of the phenomenon itself. Therefore, we cannot separate the ontological question (what is intelligence? what is cognition? what is computation?) from the epistemological question (how to explain intelligence? how to explain neural function?)

**Philosophical theories of mind:**
- Classical computational theory of mind
- Behaviourism
- Type-identity theory
- Functionalism
- Representational theory of mind
- Connectionism
- Embodied Dynamicism

**Philosophical theories of computation:**
- Formal
- Mechanistic
- Information processing
- Modeling

## Theories of Scientific Explanation and Their Weaknesses

| Theories | Criticisms and Examples |
|---|---|
| **The Deductive-Nomological (DN) Model (Hempel 1959, 1965)**<br>(aka Covering Law Model)<br>A scientific explanation is a deductive argument based on at least one "law of nature" that shows that the explanadum (the phenomenon to be explained) is to be expected given the premises of the explanans (the thing that does the explaining.)<br>• "law" is used to differentiate deterministic laws from other true generalizations that are only "accidentally true"<br>• Predictivism: "any predictively adequate model possesses explanatory force" (Kaplan 2011) | *Explanatory asymmetries*: some explanations are directional. (e.g. length of shadow cast by a flagpole)<br>  <br>*Explanatory Irrelevancies*: A derivation may satisfy the DN model, while relying on a true generalization that is irrelevant to the explanadum. E.g.<br>(L) All males who take birth control pills regularly fail to get pregnant<br>(K) John Jones is a male who has been taking birth control pills regularly<br>(E) John Jones fails to get pregnant |
| **The Statistical Relevance (SR) model (Salmon 1971)**<br>Explanations must cite causal relationships and causal relationships are captured by statistical relevance relationships, i.e. conditional dependence<br><br>P(Pregnancy\|Male.Takes birth control pills) = P(Pregnancy\|Male) = 0<br>P(Pregnancy\|Female.Takes birth control pills) ≠ P(Pregnancy\|Female) | • Objective homogeneity condition: "there are no additional omitted variables that would affect the probability"<br>• Casual relationships are greatly underdetermined by statistical relevance relationships (Cartwight, 1979 and Spirtes, Glymour and Scheines, 1993, 2000.) |
| **Functional Explanation (Cummins 1975, 1983)**<br>Phenomena are explained by reference to the functional role and organization of their subcomponents. This process of decomposing a complex capacity into simpler subcapacities can be repeated in a reductionist manner until the subcomponents are well-understood. Two stages: analysis and instantiation. | • Most easily applied when assuming functionalist/information processing theory of mind and computation. Under other conceptions of mind, it is not obvious that subcapacities should explain supercapacities.<br>• Does not establish causal relevance of subcapacities.<br>• In practice, scientists often present functional analysis as explanatory.<br>• Computational chauvinism: in line with functionalism, cognition can be explained independently of the physical implementation in the brain |
| **The Unification Model (Kitcher 1989)**<br>explanation is "a matter of unifying diverse beliefs under a few simple argument patterns"<br>"Science advances our understanding of nature by showing us how to derive descriptions of many phenomena, using the same patterns of derivation again and again, and in demonstrating this, it teaches us how to reduce the number of facts that we have to accept as ultimate." (Kitcher 1989: 423) | • The most unifying statements are not necessarily explanatory (e.g. taxonomies)<br>• No notion of causality |
| **The Causal Mechanical (CM) Model (Salmon, 1984, 1998)**<br>Constiutive (or componential) mechanistic explanation reveals the mechanisms underlying a phenomenon---the physical entities and activities organized such that they exhibit the explanadum---and thereby shows how a phenomenon is produced by its causes. Organization is important: not just the sum of parts but their causal interaction.<br><br>Aspects of mechanistic explanation:<br>1. the nature of the phenomenon to be explained<br>2. the constitutive relationship between a phenomenon and its components<br>3. the difference between real components and useful fictions<br>4. the nature of capacities or activities<br>5. the nature of mechanistic organization<br>6. the nature of constitutive explanatory relevance | Not clear how the CM model would apply to higher-level, more abstract phenomena, e.g. cognition.<br>Physical components may be irrelevant to explain some phenomena<br>  <br><br>  <br> |
| **Simulation as explanation**<br>Scientists who develop simulations often claim that simulations represent or demonstrate progress towards the goal of explaining the simulated phenomenon. | Philosophers seems to agree that simulations are, in general, not explanatory. From a causal perspective, identifying one "how possibly" mechanism out of potentially infinitely many, does not in itself constitute progress towards finding the true mechanism. However, Grune-Yanoff (2009) suggests that simulations may be seen as candidate functional explanations. |

## Three Perspectives in Cog Comp Neuro

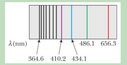
### What neuroscientists say

**Ilya Nemenman: "It doesn't matter if it's true"**

Good theories/models of complex biological systems:
- are phenomenological
- make accurate predictions
- throw away unnecessary details
- are as simple as possible (Occam's Razor)
- explain a limited set oh phenomena
- are falsifiable but not falsified, according to Bayesian statistics

*The Balmer Formula*

$$\frac{1}{\lambda} = R_H\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right)$$




**Kendrick Kay: Explain via functional analysis**

"Mental operations [...] can be viewed as information processing operations. The cognitive neuroscientist asks: for a given brain region, what stimulus, cognitive, or motor operations are performed by neurons in that region? [...] Models posit that specific variables relate to neural activity. As such, models provide explanations of measurements of the brain [...] With appropriate experimental measurements, we can adjudicate different models and decide which model is most accurate"
To explain, compare interpretable models with relatively small number of parameters. Models are falsified to the extent that they fail to make accurate predictions.

**Jonas Kubilis: "Predict then simplify"**

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? " (Kelly, 2007)
- Build deep network-based models that predict neural activity as well as possible for the broadest set of experiments/stimuli
- Then narrow to small number of principles rather than parameters that are integral to the system
- As our models demonstrate increasingly realistic behaviours and mimic neural representations with increasing fidelity, we'll understand the system better

### What philosophers say

Such models are certainly useful, but are they explanatory?
- Statistical relevance model
- Predictivism

- Cognitive neuroscience explains cognitive phenomena via functional analysis of the brain
- Information processing theory of mind
- Selective response of subcomponent is indicative of its role in the capacity to be explained.

- Predictivism
- Unification model
- Simulation
- Abandons commitment to information processing theory of mind (or any theory of mind?)
- Imprecise about how turning the crank between better models and more naturalistic experiments will ultimately lead to unifying explanation

## Looking Forward

**Constraints on explanations (according to Craver 2007)**
- Mere temporal sequences are not explanations
- Causes explain effects and not vice versa
- Causally independent effects of common causes to not explain one another
- Causally irrelevant phenomena are not explanatory
- Causes need not make effects probably to explain them

**Artificial neuroscience and biological machine learning**
Swapping methodologies, approaches and philosophies 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.
- Empirical analyses of deep learning systems that seem almost neuroscientific (e.g. ablation analyses, receptive field analysis, psychophysics) but with reference to concepts from deep learning theory (e.g. generalization, expressivity).
- Machine learning inspired neuroscience

**Desiderata for a new theory of explanation for both artificial and biological intelligence:**
- Reflect that learning is central to intelligence
- Multiple realizability without computational chauvinism
- Abandon focus on physical computation
- Not concerned with characterizing the specific function that is computed by a network
  1. because we probably can't glean anything meaningful from that anyway (see "Can neural computation be compressed enough for us to understand it?" Lillicrap & Kording, this meeting)
  2. because we know that repeated optimizations of the same network lead to solutions that occupy distinct regions in function space, many local minima (Erhan 2010)
- Take the good parts from existing theories: causality (CM model), unifying principles (unification model), abstraction (functional)

## Conclusion

- 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.
- Let's be precise with our use of the words "explanation", "mechanism", "description". Not a value judgment—description is important too!
- Room for increased clarity about what phenomena we are studying, what our short term and long term goals are, and how our short term goals will serve the ultimate goal of explaining the phenomenon in question. A we studying visual object recognition or the neural activity elicited during visual object recognition?

## References


[1] Nemenman, I. (2018). Playing Newton: Automatic Construction of Phenomenological, Data-Driven Theories and Models. https://simons.berkeley.edu/talks/ilya-nemenman-4-17-18.
[2] Kaplan, D. M. (2011). Explanation and description in computational neuroscience. Synthese, 183 (3).
[3] Kay, K. N. (2017). Principles for models of neural information processing. NeuroImage.
[4] Kubilius, J. (2017). Predict, then simplify. NeuroImage.
[5] Woodward, J. (2017). Scientific Explanation. In E. N. Zalta (Ed.), The stanford encyclopedia of philosophy (https://pled.). Metaphysics Research Lab, Stanford University.
[6] Kelly, K. T. (2007). Simplicity , Truth , and the Unending Game of Science. In S. Bold, Benedikt Lowe, T. Rasch, & J. van Benthem (Eds.), Foundations of the formal sciences v: Infinite games
[7] Grüne-Yanoff, T., & Weirich, P. (2010). The philosophy and epistemology of simulation: A review. Simulation and Gaming, 41(1), 20–50.