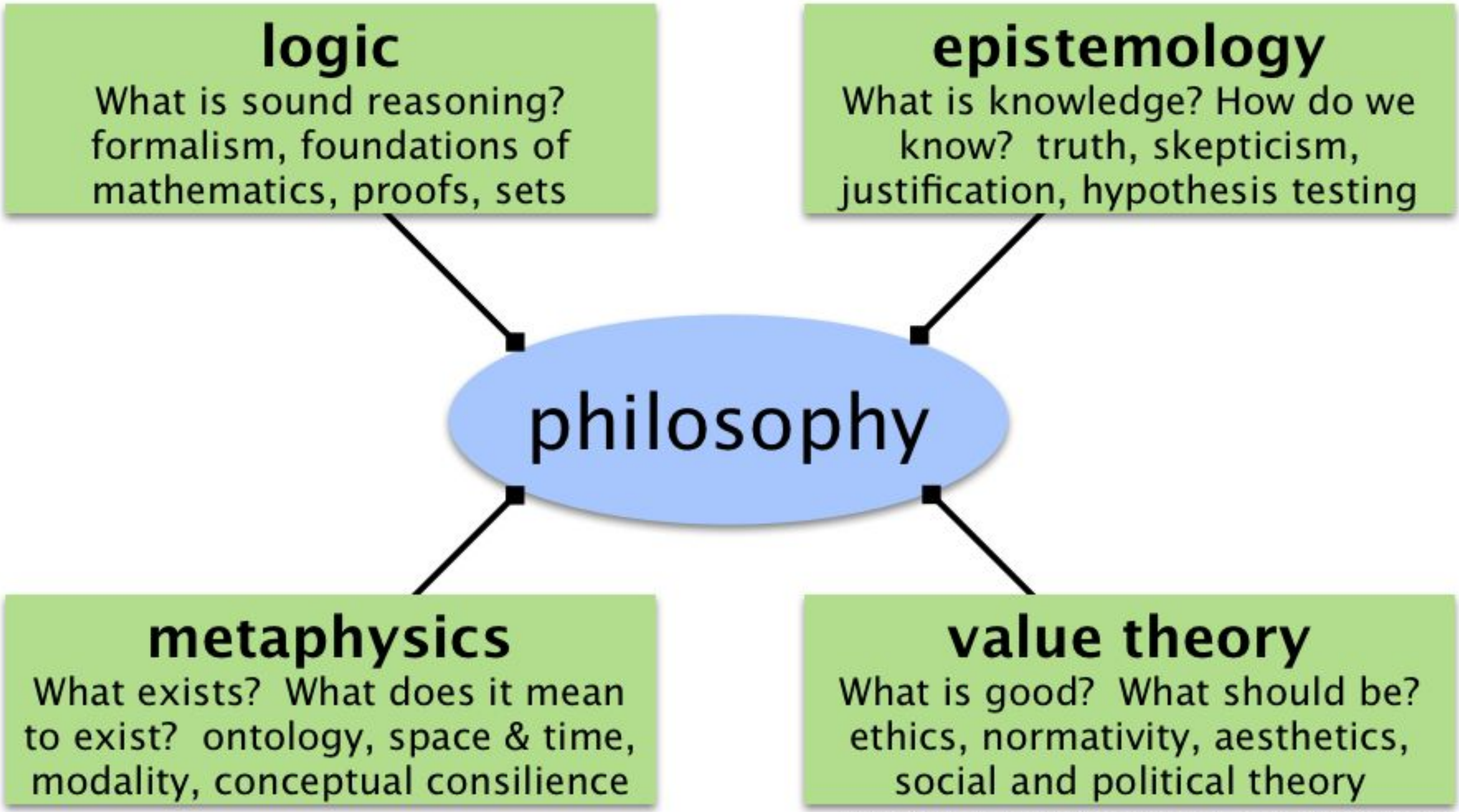
Theories of explanation in artificial and biological intelligence

Jessica Thompson

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 roadmap for how to make progress towards at least one scientific goal
- When the goal is to explain, this roadmap requires some idea of what type of explanation we're looking for.
- Claim 1: Science of intelligence weak in this domain, little agreement or formalization of what would be acceptable explanations of the phenomena of intelligence
- Claim 2: Integration of deep learning and neuroscience will require a common theory of explanation that applies to both artificial and biological intelligence



philosophy

logic

What is sound reasoning? formalism, foundations of mathematics, proofs, sets

epistemology

What is knowledge? How do we know? truth, skepticism, justification, hypothesis testing

metaphysics

What exists? What does it mean to exist? ontology, space & time, modality, conceptual consilience

value theory

What is good? What should be? ethics, normativity, aesthetics, social and political theory

Outline

- The epistemological question: Theories of scientific explanation
- The ontological question: Theories of mind and computation
- Descriptive analysis of what computational neuroscientists say about their philosophical commitments
    - How do scientists invoke the aforementioned theories of explanation, mind and computation?
- Discuss challenges with trying to construct a common theory of explanation for biological and artificial intelligence

Theories of Scientific Explanation

- The Deductive-Nomological (DM) Model (Hempel 1965)
- The Statistical Relevance (SR) model (Salmon 1971)
- The Unification Model (Kitcher 1989)
- The Causal Mechanical (CM) Model (Salmon, 1984, 1998)

The role of a theory of scientific explanation

1. Characterize the structure of explanations in science
2. Distinguish between explanations that are scientific and those that are not
3. 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.
   - **phenomenological models** (aka “descriptive model”): 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.

Criteria of adequacy for an account of explanation (according to Craver):

1. **Descriptively adequate**: does it match reality or only some ideal?
2. **Demarcate** explanation from other types of scientific achievements e.g. categorization, simulation
   - "For example, an account of explanation should make sense of the difference between simulating or modeling a phenomenon and explaining it. Ptolemaic models can be used to simulate and predict planetary motion across the night sky but they do not explain it; the epicycles, deferents, and equants are merely mathematical tools in the models with no basis in the structure of the heavens. An explanation, in contrast, shows why the planets move as they do and allows one to say how they would move if conditions were different." (p. 20 Craver 2007)
3. Reveal criteria for **assessing** explanations

The Deductive-Nomological (DN) Model (Hempel 1965)

aka Hempel's model, the Hempel–Oppenheim model, the Popper–Hempel model, or the covering law (CL) model

- Scientific explanation consists of an
  - **explanadum**: the thing to be explained
  - **explanans:** the thing that does the explaining, but only if several conditions are met:
    1. “the explanandum must be a logical consequence of the explanans”. An explanation is the **deductive argument** that shows that the explanadum is **expected** given the premises of the explanans.
- The explanans must rely on at least one "**law of nature**" in its explanatory logic.
  - "law" is used to differentiate deterministic laws from other true generalizations that are only "accidentally true"
- Predictivism: “any predictively adequate model possesses explanatory force” (Kaplan 2011)

Woodward, James, "Scientific Explanation", The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/>

The Deductive-Nomological (DM) Model

There are a number of well known counterexamples to the claims that the DN model provides necessary or sufficient conditions for explanation.

- **Explanatory asymmetries**: derivation of an explanadum from a law and initial conditions can meet the criteria for a DN explanation, while the reverse derivation of initial conditions from the explanadum and law is not explanatory. The DN model doesn't account for the fact that some explanations are directional. (e.g. length of shadow cast by a flagpole)
- **Explanatory Irrelevancies**: A derivation may satisfy the DN model, while relying on a true generalization that is irrelevant to the explanadum. E.g.
  - (L) All males who take birth control pills regularly fail to get pregnant
  - (K) John Jones is a male who has been taking birth control pills regularly
  - (E) John Jones fails to get pregnant

Woodward, James, "Scientific Explanation", The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/>

The Statistical Relevance (SR) model (Salmon 1971)

motivation for the SR model:

1. Explanations must cite causal relationships
2. Causal relationships are captured by statistical relevance relationships

$$P(\text{Pregnancy} \mid \text{T.Male.Takes birth control pills}) = P(\text{Pregnancy} \mid \text{T.Male}) = 0$$

$$P(\text{Pregnancy} \mid \text{T.Female.Takes birth control pills}) \neq P(\text{Pregnancy} \mid \text{T.Female})$$

problems:

- objective homogeneity condition: “there are no additional omitted variables that would affect the probability”
- 2. is false, casual relationships are greatly underdetermined by statistical relevance relationships (Cartwight, 1979 and Spirtes, Glymour and Scheines, 1993, 2000.)

Woodward, James, "Scientific Explanation", The Stanford Encyclopedia of Philosophy (Fall 2017 Edition), Edward N. Zalta (ed.), <https://plato.stanford.edu/archives/fall2017/entries/scientific-explanation/>

The Unification Model (Kitcher 1989)

explanation is “a matter of unifying diverse beliefs under ar few simple *argument patterns*”

“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)

problem: nonexplanatory unification (e.g. taxonomies) still missing causality

The Causal Mechanical (CM) Model (Salmon, 1984, 1998)

- “mechanisms are entities and activities organized such that they exhibit the explanandum” (Craver 2007)
- explanation is "a matter of showing how a phenomenon is produced by its causes", i.e. of situating a phenomenon in the causal structure of the world
- two types:
    - etiological: explanation in terms of antecedent causes, e.g. virus causes flu, dehydration causes thirst
    - constitutive (or componential): describe underlying mechanisms

Norms of constitutive mechanistic explanation

- Reductive tradition
  - understanding is the rational expectation of a phenomenon at one level from laws governing parts at a lower level
- Systems tradition
  - explanation is the matter of decomposing a system into its parts and demonstrating how those parts are organized in such a manner that they exhibit the phenomenon to be explained.
  - “If you can’t make one, you don’t know how it works”
  - “you need a blueprint, a recipe, an instruction manual, a program” (Dretske 1994: 468)
  - Organization is important:
    - not just the sum of parts but their interaction
  - separate how-possibly from how-actually

Aspects of mechanistic explanation (Craver 2007)

1. the nature of the phenomenon to be explained
2. the constitutive relationship between a phenomenon and its components
3. the difference between real components and useful fictions
4. the nature of capacities or activities
5. the nature of mechanistic organization
6. the nature of constitutive explanatory relevance

Can cognitive phenomena be explained mechanistically?

- Requirement that components be physical
- Abstraction allowed but how?

Theories of Mind and Computation

Should we think of the brain as a computing system?

What distinguishes computing systems from non-computing systems?

Theories of mind

- Classical computational theory of mind
- Behaviourism
- Type-identity theory
- **Functionalism**
- Representation theory of mind
- Connectionism
- Embodied Dynamicism

Theory of computation

- Formal
- Mechanistic
- **Information Processing**
- Modelling

What about learning? Historically not centered in theories of mind nor computation

What do neuroscientists say?

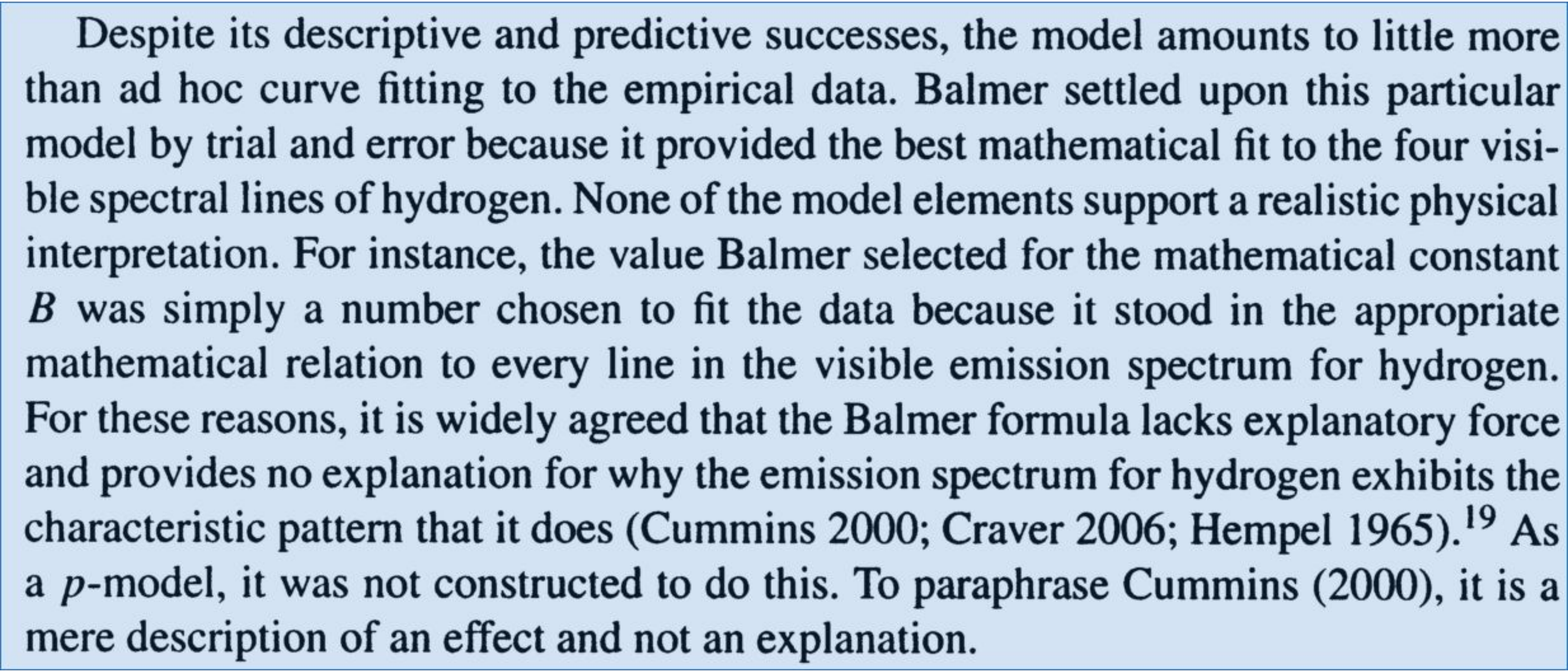
Ilya Nemenman: It doesn’t matter if it’s true

- good theories/models of complex biological systems
  - are phenomenological
  - make accurate predictions
  - adhere to Occam’s razor
- but what are they good *for?*
- Balmer formula example

Consider the Balmer formula (Eq. 1), which Balmer constructed to describe the four visible lines in the emission spectrum of hydrogen:

$$\lambda = B\left(\frac{m^2}{m^2 - n^2}\right) \tag{1}$$

where λ is the wavelength, B is a constant with the value of 364.56 nm, n equals 2, and m is variable taking integer values greater than 2 ($m > n$). When m is replaced by any integer from the appropriate range, the model outputs the wavelength for a line in the emission spectrum of hydrogen. The Balmer formula is descriptively adequate. It provides an accurate mathematical description of the target phenomenon and can thus be said to "save" it. The model is also useful for making predictions. Although initially constructed to account for the visible spectral lines for hydrogen, the model successfully predicted the existence of several unobserved spectral lines outside the visible range that were later experimentally confirmed.



Despite its descriptive and predictive successes, the model amounts to little more than ad hoc curve fitting to the empirical data. Balmer settled upon this particular model by trial and error because it provided the best mathematical fit to the four visible spectral lines of hydrogen. None of the model elements support a realistic physical interpretation. For instance, the value Balmer selected for the mathematical constant *B* was simply a number chosen to fit the data because it stood in the appropriate mathematical relation to every line in the visible emission spectrum for hydrogen. For these reasons, it is widely agreed that the Balmer formula lacks explanatory force and provides no explanation for why the emission spectrum for hydrogen exhibits the characteristic pattern that it does (Cummins 2000; Craver 2006; Hempel 1965).[19] As a *p*-model, it was not constructed to do this. To paraphrase Cummins (2000), it is a mere description of an effect and not an explanation.

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

Kendrick Kay: explain via functional analysis

- Cognitive neuroscience: “The question of understanding how the functions of the physical brain can yield the thoughts and ideas of an intangible mind” (Gazzaniga et al., 2014).
- “It is widely accepted that “thoughts and ideas of an intangible mind,”or mental operations more generally, 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?”

Kay, K. N. (2017). Principles for models of neural information processing. NeuroImage

Kay on explanation

“Models posit that specific variables relate to neural activity. As such, models provide explanations of measurements of the brain. For example, suppose we find that a neuron is highly active when a clip of rock music is played but is only weakly active when a speech clip is played. Why does this occur? One model could be that the neuron computes overall sound intensity, and the reason we observe weak activity for the speech clip is that it has low sound intensity. Alternatively, there are other candidate models that might explain the phenomenon (e.g., selectivity for guitar tones, variations in attentional engagement). With appropriate experimental measurements, we can adjudicate different models and decide which model is most accurate (Naselaris and Kay, 2015).”

Kay, K. N. (2017). Principles for models of neural information processing. NeuroImage

Kay’s account of explaining neural systems

When you assume the brain is an information processing system, then functional analysis provides a good explanation for animal behaviour/perception.

”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).

The behaviour of a system is explained by identifying and labeling the different signals it contains/computes at different locations.

Under a different theory of mind/computation, this account of explanation breaks down.

Jonas Kubilius: 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? You don’t have to be a card-carrying skeptic to wonder what the tacit connection between simplicity and truthfinding could possibly be” (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
- The predictivist gap: eventually if we get better and better at making predictions, we’ll have to understand the system better (not sure how)

- Kubilius, J. (2017). Predict, then simplify. *NeuroImage*
- 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.

Constraints on explanations (Craver 2007)

1. mere temporal sequences are not explanations
2. causes explain effects and not vice versa
3. causally independent effects of common causes to not explain one another
4. causally irrelevant phenomena are not explanatory
5. causes need not make effects probably to explain them

Craver, Carl, Explaining the Brain, Oxford University Press, 2007

Desiderata for a theory of explanation for artificial and biological intelligence

- Learning as central to intelligence
- Multiple realizability without computational chauvinism
- Abandon focus on physical computation
- Not concerned with describing a specific function that is computed
-

Conclusion

- No satisfactory theories of explanation for artificial and/or biological intelligence
- No expectation to even appeal to one to justify one’s approach
- Description != explanation
- Most of neuroscience seems to be hunch following without any formalized idea about how to compose and validate explanations
- Machine learning is using empirical approaches that mimic analysis of neuro data, exciting opportunity to integrate with empirical neuroscience