# NACS 645 –
# Is AI part of Cognitive Science?

-

Valentin Guigon


DEPARTMENT OF
PSYCHOLOGY



PROGRAM IN
NEUROSCIENCE &
COGNITIVE SCIENCE

# Information processing

- 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)
- *Thoughts and ideas of an intangible mind means information-processing operations* (representing info, storing info, reasoning about the info, acting from the info)
- Information processing:
  - A (cognitivist) framing in which cognitive systems are described as transforming structured inputs into structured outputs (i.e., computational transformations).
- Mental operations:
  - Computational transformations defined over representational states (i.e., configuration of the system that carries information about something)

Do DNNs have merit as models of biological systems?

# Models in cognitive neuroscience: goals and levels

Kay, 2018. *NeuroImage*.

A **model** is a formal or quantitative specification of how a system produces its observable outputs from its inputs.

## Types of models (by form)

- **Qualitative models**
  Verbal or conceptual descriptions with loosely specified assumptions
- **Quantitative models**
  Explicit, parameterized mappings enabling falsifiable predictions
- **Phenomenological models**
  Fit data without specifying underlying mechanisms
- **Oracle models**
  Rely on externally supplied or human-labeled features; can predict neural

## Levels of explanation (Marr)

- **Functional models (computational / algorithmic level)**
  Specify the transformation from inputs to outputs performed by a neuron or population of neurons, without committing to biological implementation.
- **Mechanistic (biophysical / circuit) models (implementational level)**
  Describe how biological components (neurons, circuits, dynamics) implement those transformations.

*A model can be successful at one level without satisfying the other.*

## Goals of models (by scientific role)

- **Descriptive**
  Compress complex, noisy data into a small number of interpretable parameters, enabling comparison across conditions or systems.
- **Explanatory**
  Identify variables or mechanisms that account for why observed activity has the form it does.
- **Predictive**
  Generate testable predictions of system behavior under conditions not used to fit the model.

# Philosophical commitments behind models of cognition

Perconti & Plebe, 2020. *Cognition*.

Whether a model is considered *legitimate* in cognitive science depends on underlying assumptions about **how cognitive structure arises**.

## Connectionism (cognitive science)

- A family of cognitive theories in which mental processes arise from the interaction of many simple units connected in parallel
- Knowledge is encoded in distributed patterns of activation and connection weights
- Cognitive functions emerge through learning-driven changes in these systems

→ Supports **learning-based**, non-symbolic models (e.g., neural networks)

## Empiricism

- **Empiricism (philosophy)**
  - The view that concepts and knowledge derive primarily from sensory experience
- **Empiricism (cognitive science)**
  - Cognitive architectures acquire structure through exposure to data rather than via innate symbolic rules
  - Connectionist and deep learning models instantiate this position

→ Emphasizes **learning, exposure, and experience** as sources of structure

## Rationalism

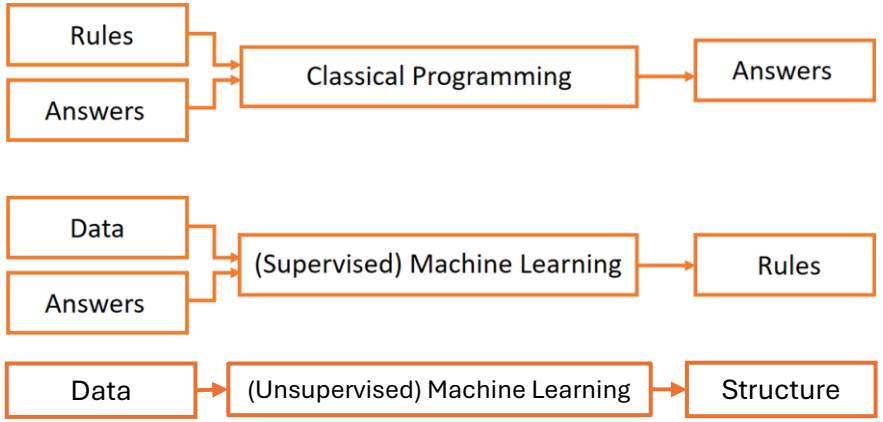
- **Rationalism (philosophy)**
  - The view that some concepts, structures, or inferential rules are innate or accessible a priori
- **Rationalism (cognitive science)**
  - Cognitive architecture includes innate representational primitives and rules
  - Knowledge is encoded in structured, rule-based symbolic formats; cognition is modeled as syntactic manipulation of symbols

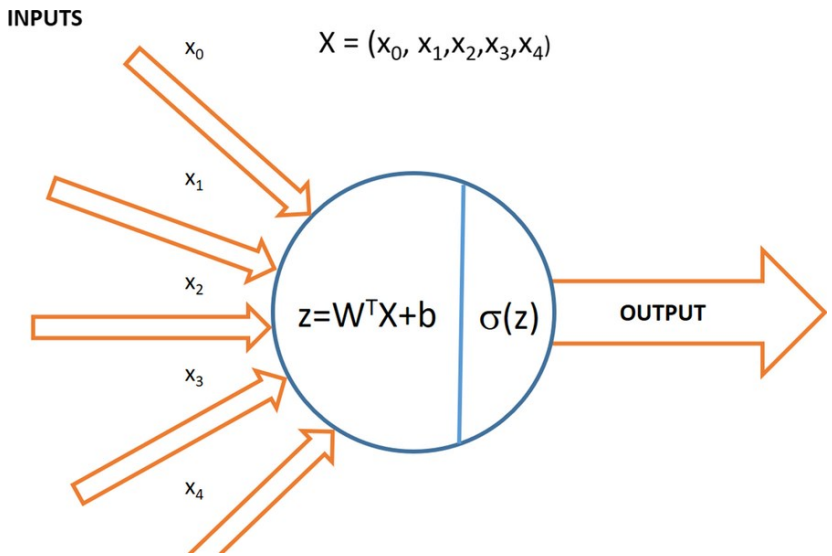
→ Grounds the **symbolic tradition**

# Artificial neurons

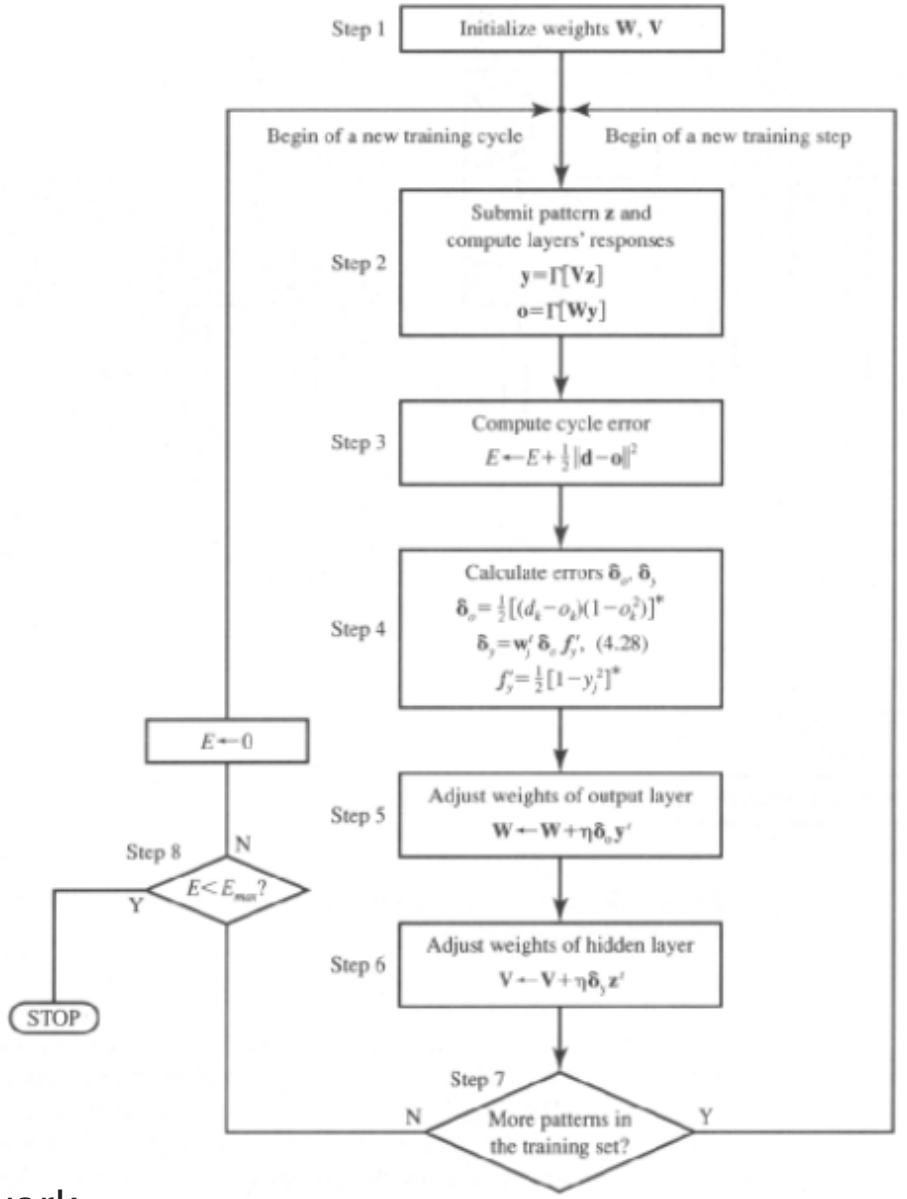
Iglesias et al., 2021. *Insights into Imaging.*
N.P. Rougier. M2 Cognitive Science. *University of Bordeaux*

## Classical programming vs machine learning



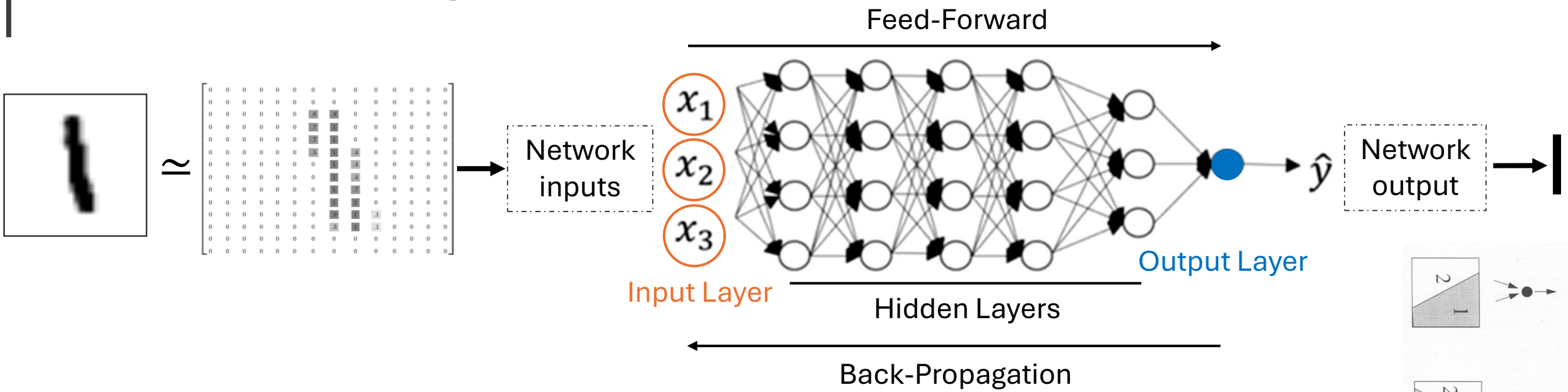

Structure of a single neuron in an artificial neural network

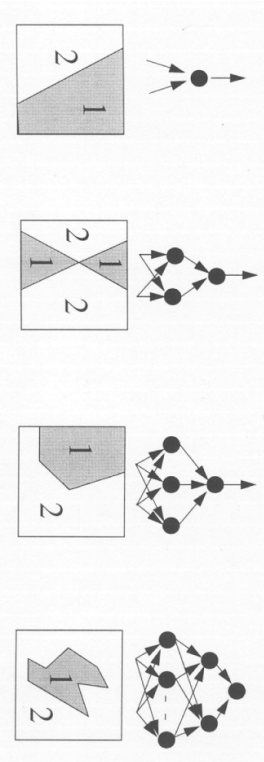



1. Initialization
2. Present training example (forward pass)
3. Compute global error (loss)
4. Compute local error signals (gradients)
5. Update weights in hidden layers
6. Update weights in input-connected layers
7. Repeat

Iglesias et al., 2021. *Insights into Imaging*.

# Deep Learning



**Lingo**

- **Perceptron**:
  Early single-layer *linear* classifier (limited to linearly separable problems)
- **Feedforward architecture**:
  Layered mapping from input to output
- **Backpropagation**:
  Gradient-based algorithm for adjusting weights with chain rule
- **Stochastic gradient descent**:
  Mini-batch approximation enabling efficient learning in high dimensions
- **Deep vs shallow networks**:
  Depth refers to the number of hidden layers; greater depth increases representational expressivity but complicates optimization

Brette, 2019. *Behavioral and Brain Sciences.*
*Mathis et al., 2024.* Cell.

# Are DNNs models of cognition?

A *model of cognition* may refer to: models of **behavior** & models of **internal computations** (functional models), or models of **biological mechanisms** (mechanistic models).

## Criteria for evaluating models

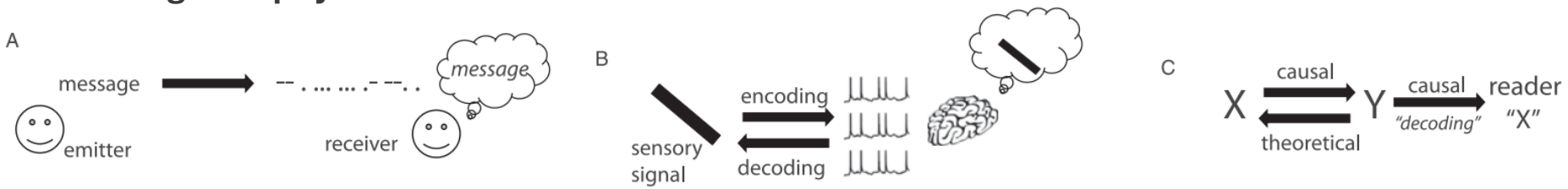
- A system can be **behaviorally successful** without being a good **model of cognition**.
- A system can be **cognitively informative** without being **biologically realistic**.
- Good models are expected to achieve:
  - **Accuracy:** prediction of held-out data; generalization across stimuli, tasks, etc.
  - **Understanding:** clarity about internal components, assumptions, operations
  - May also require to respect known facts about **cortical org. and physio.**

## Encoding in contemporary neuroscience

- **Neural populations** are treated as **encoding internal variables** (sensory features, latent task variables, beliefs, values, motor plans) **with neural responses**
- **Upstream populations** encode **variables**; **downstream populations** decode **population activity** to infer task-relevant variables or guide behavior
- Linearity vs non-linearity is a **modeling choice**, dependent on brain region, task demands, etc. Multiple model classes are used (often in combination): **Linear models**, **Non-linear models**, **DNN**

## Why deep learning enters neuroscience

- **Simpler models often fail to capture complex stimulus structure**, hierarchical feature transformations, distributed population codes
- DNNs provide powerful **non-linear encoding models**, effective **decoders of neural population activity**, learned latent representations aligned with neural data, high predictive accuracy for neural and behavioral responses




The coding metaphor

*Mathis et al., 2024.* Cell.

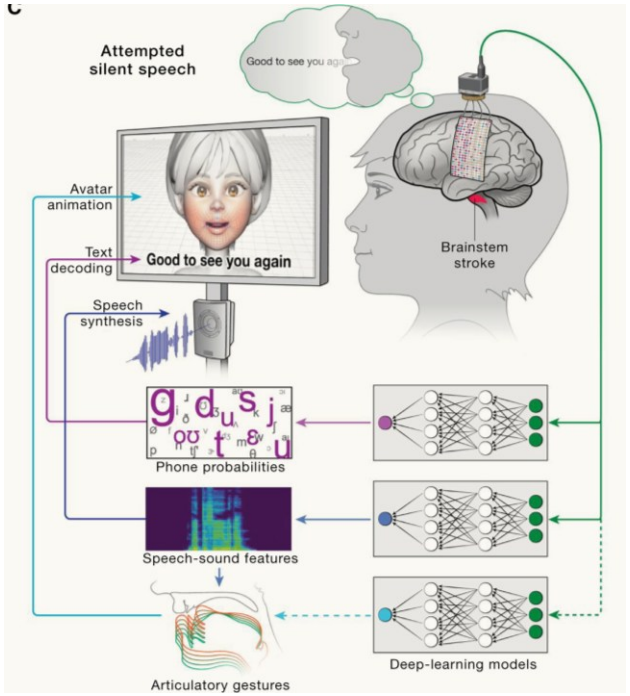
# What DNNs contribute

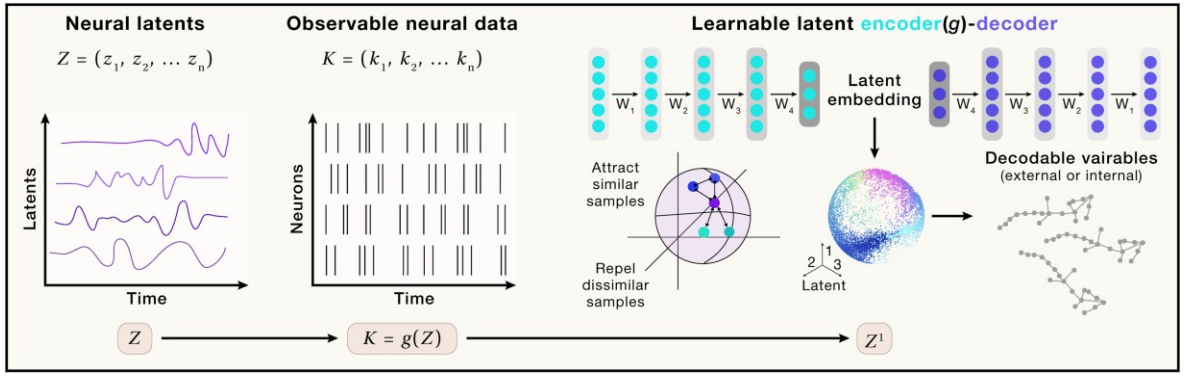
## What DNNs demonstrably provide

- High predictive accuracy in vision and language
- Learned internal representations that correlate with neural population activity and capture task-relevant latent variables
- Flexible *in silico* systems

## What they do not automatically provide

- Biological realism at the cellular or circuit level
- Embodiment, development, or ecological coupling
- Direct mechanistic explanations of neural implementation




## 4E cognition: what is being challenged

- **Embodied:** cognition depends on bodily constraints
- **Embedded:** cognition exploits environmental structure
- **Enactive:** cognition is constituted by action, not internal representation alone
- **Extended:** cognitive processes may span brain, body, and environment

## Is representational modeling present and necessary?

- DNNs succeed without bodies, without sensorimotor loops, without ecological interaction
  - Yet they exhibit internal representations predictive of neural data

# NACS 645 – Cognitive properties of LLMs

-

Valentin Guigon


DEPARTMENT OF
PSYCHOLOGY



PROGRAM IN
NEUROSCIENCE &
COGNITIVE SCIENCE

# Sentiments towards cognitive properties of LLMs

## Incentive-driven stances

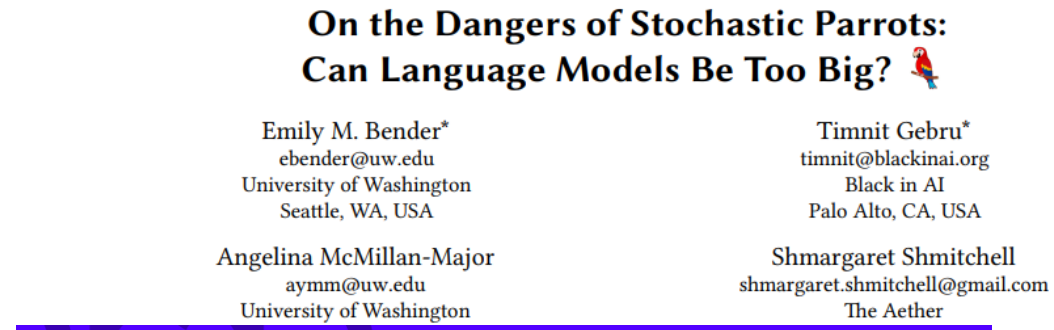
- **Corporate actors**: hype inflation
- **VCs**: AGI inevitability
- **Media**: oscillation between miracle and existential threat
- **Public users**: anthropomorphism, emotional attachment

## Paradigm-rooted stances

- **Chomskyans**: limits of statistical learning
- **Connectionists:** prospects of statistical learning
- **Classical cognitivism**: LLM architectures reduces cognition to a single objective function (next-token prediction)
- **RL-centric & world-model views**: language models as dead ends (no learning; no physical properties)
- **Bayesian brain**: Next-token prediction ≠ predictive coding
- Plus deflationary claims that protect threatened existing frameworks

## Psychological reactions

- Mesmerized engineers
- Optimistic researchers
- Pessimistic researchers
- Ethical skeptics
- …

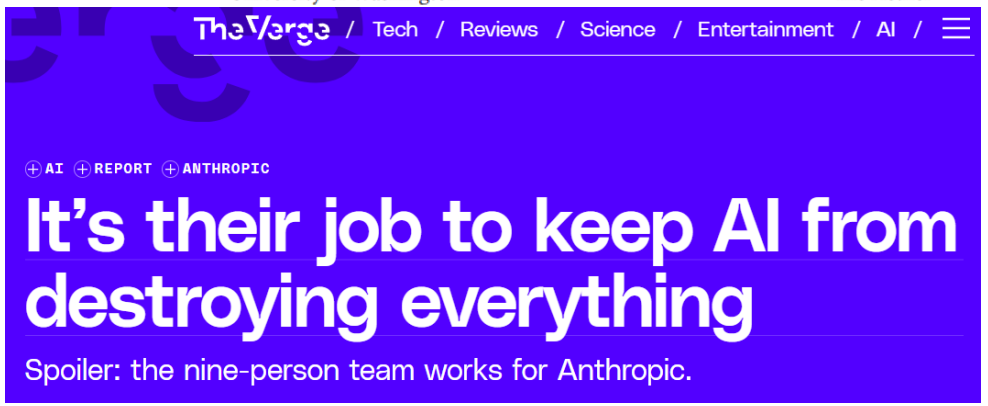

On the Dangers of Stochastic Parrots:
Can Language Models Be Too Big?

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether






Meta chief AI scientist Yann LeCun is leaving to create his own startup

PUBLISHED WED, NOV 19 2025•4:31 PM EST | UPDATED WED, NOV 19 2025•8:58 PM EST

Jonathan Vanian
@IN/JONATHAN-VANIAN-B704432/

SHARE

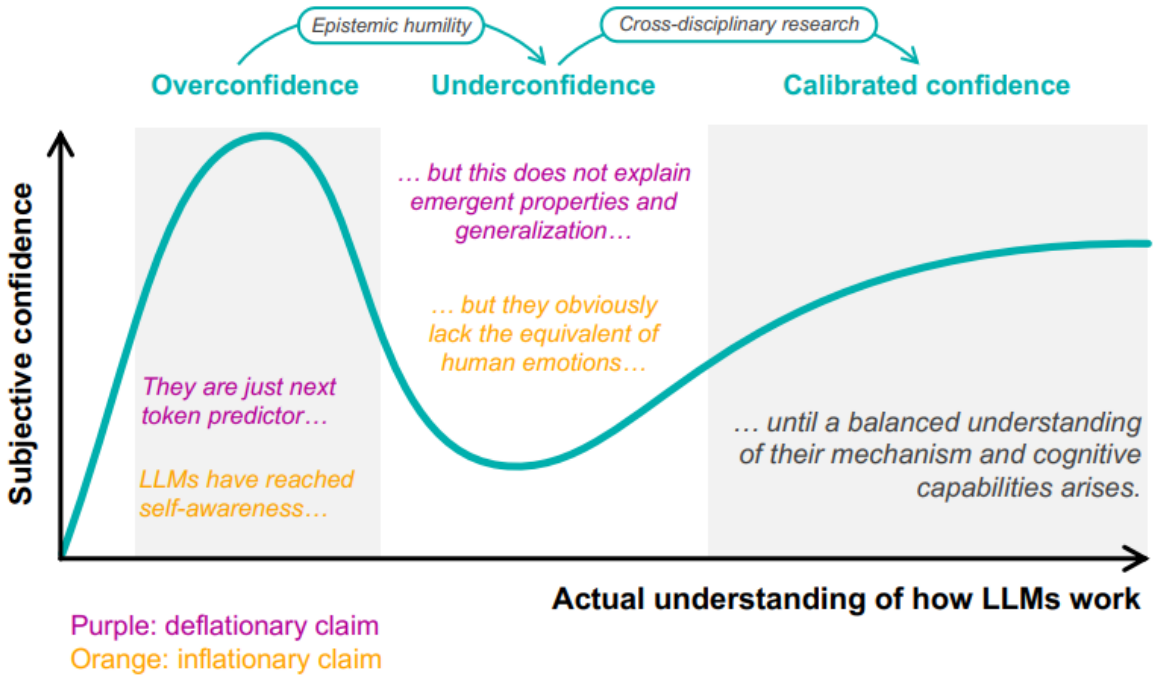
Palminteri & Pistilli, 2025. *arXiv*.

# The challenge of interpreting LLMs output

## Goal: reaching open-minded skepticism and epistemic humility

| Cognitive bias | Inflationary claim | Deflationary claim |
|---|---|---|
| Cognitive dissonance | *I've developed an emotional connection to an AI assistant, therefore it must have genuine sentience.* | *I've dedicated my career to expert systems and traditional AI approaches, so Deep Learning cannot be truly intelligent.* |
| Wishful thinking | *I sympathize with the transhumanist movement, therefore the singularity is near.* | *I am anxious about potential AI risks, therefore AI is still far from Artificial General Intelligence.* |
| Illusion of depth of understanding | *I devised a new fine-tuning methods for LLMs, so they are now capable of advanced reasoning.* | *LLMs are just trained to predict words, therefore they cannot be truly intelligent.* |

| Claim | Epistemological problem | Ethical problem |
|---|---|---|
| Inflationary | *Prematurely concluding AI has reached human-like intelligence, potentially stifling critical research and alternative approaches.* | *Risking unwarranted moral attribution, creating false expectations, and potentially manipulating human emotional vulnerabilities.* |
| Deflationary | *Missed opportunity to use LLMs to understand genuine elements of intelligence emerging in these systems.* | *Failing to anticipate and prepare for the transformative potential of AI technologies, potentially undermining necessary ethical safeguards and responsible development.* |




*" Regardless of whether these systems possess consciousness or genuine understanding, the human emotional responses and relationships they evoke are undeniably authentic and ethically significant"*

Palminteri & Pistilli, 2025. *arXiv*.

# Navigating inflationary and deflationary claims

*“Experts' clear understanding of the training objective (predicting the next token) seems to wrongly extend to a belief that they also understand the underlying mechanisms by which LLMs achieve this goal.“*

*“Understanding [LLMs simple objective]*

*[a)] does not guarantee that we truly understand how the task is achieved,*

*[b)] nor does it rule out the possibility that the strategies necessary to accomplish this goal require the development of higher-level cognitive capacities.”*

*“LLM development [...] should be scrutinized by independent researchers and regulatory bodies who can maintain scientific integrity through rigorous and transparent evaluation free from market pressures.*

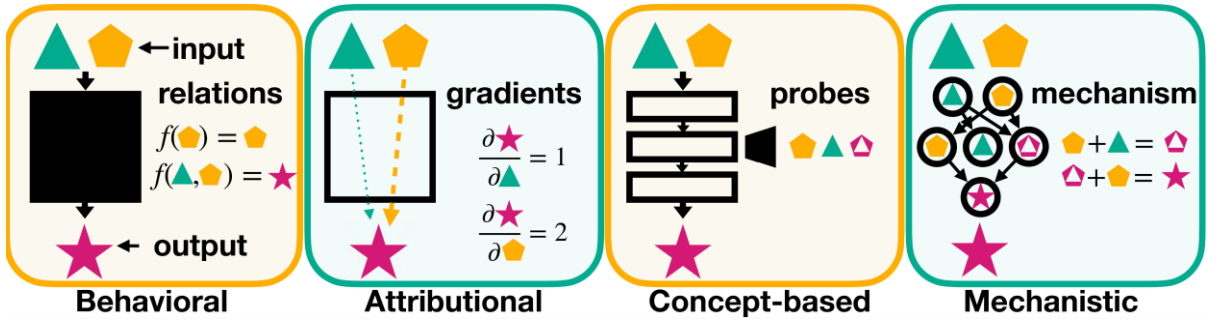
*[...]*

*When computer scientists, linguists, cognitive scientists, and philosophers work together with epistemic humility, they will create good conditions for more balanced evaluations”*

# What is happening, and what does it amount to?

## **Mechanistic Interpretability** (MechInterp)

- Goal: **reverse-engineer neural networks** by mapping how information flows through neurons, attention heads, and circuits, and identifying which components implement which computations
- **Two approaches:**
  - **Bottom-up (mechinterp proper):** uncover circuits and computations directly from learned parameters
  - **Top-down (representation engineering):** making the model's internals interpretable by design, by shaping them to match human concepts

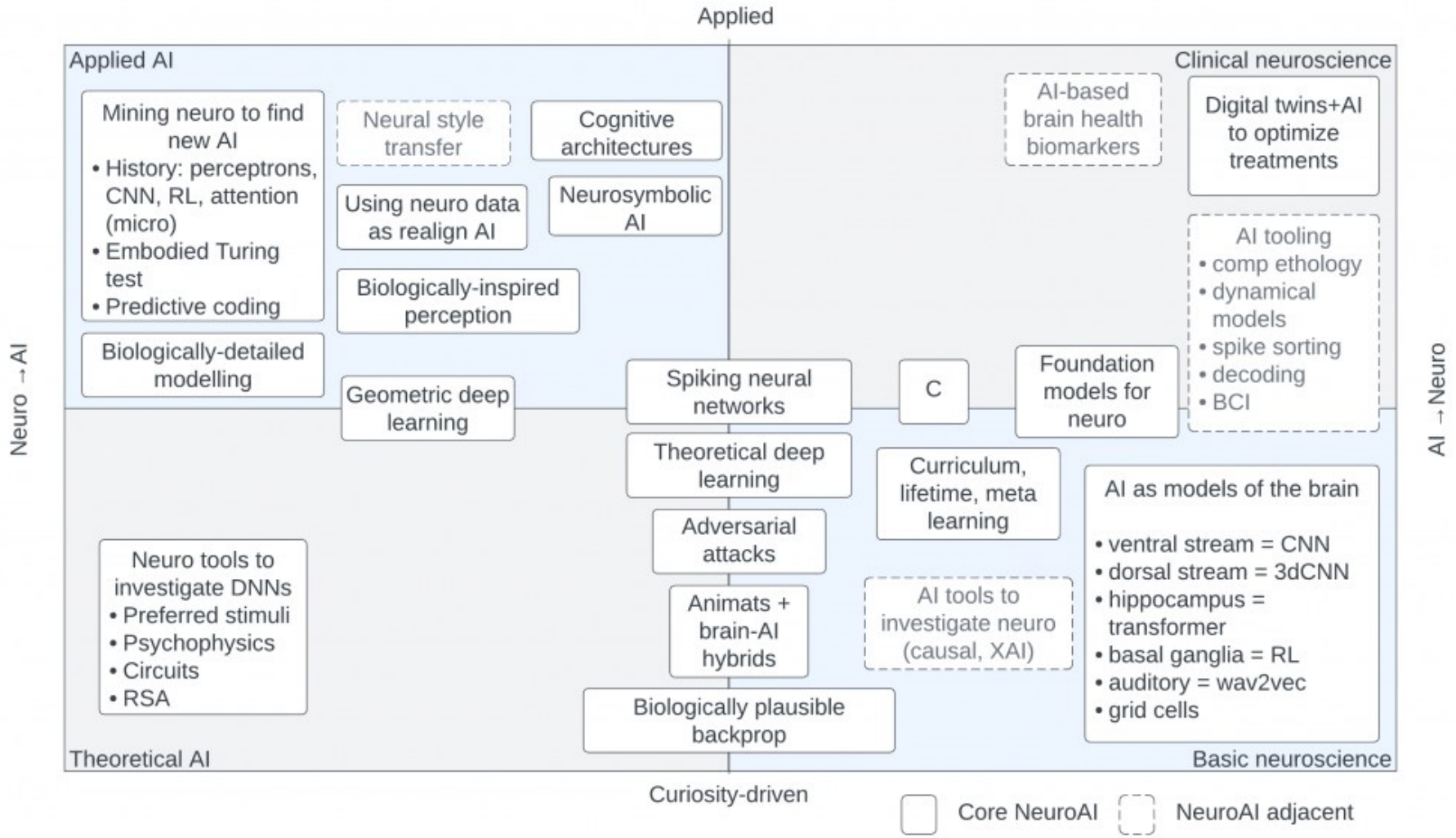



## Cognitive Interpretation

- Do a model's internal computations play roles analogous to cognitive representations or processes?
  - Low-dimensional latent structure aligned with meaningful quantities (objects, relations, goals)
  - Explicit representations, decodable with simple readouts
  - Functional use in prediction, inference, control, etc.
  - Stable and abstract representations, reusable across contexts and across processing stages
- These make states eligible for cognitive interpretation. When do such alignments license a cognitive claim and not a metaphor? How do they fit broader cognitive paradigms (e.g., connectionism, representational theories, predictive processing, 4E cognition)?
- Route not yet formalized as a single framework, but emerges as *Psychology x Linguistics x Neuroscience x Computer Science x Philosophy of Cognition* to assess what model's computations amount to cognitively

# From Neuroscience to AI and AI to Neuroscience

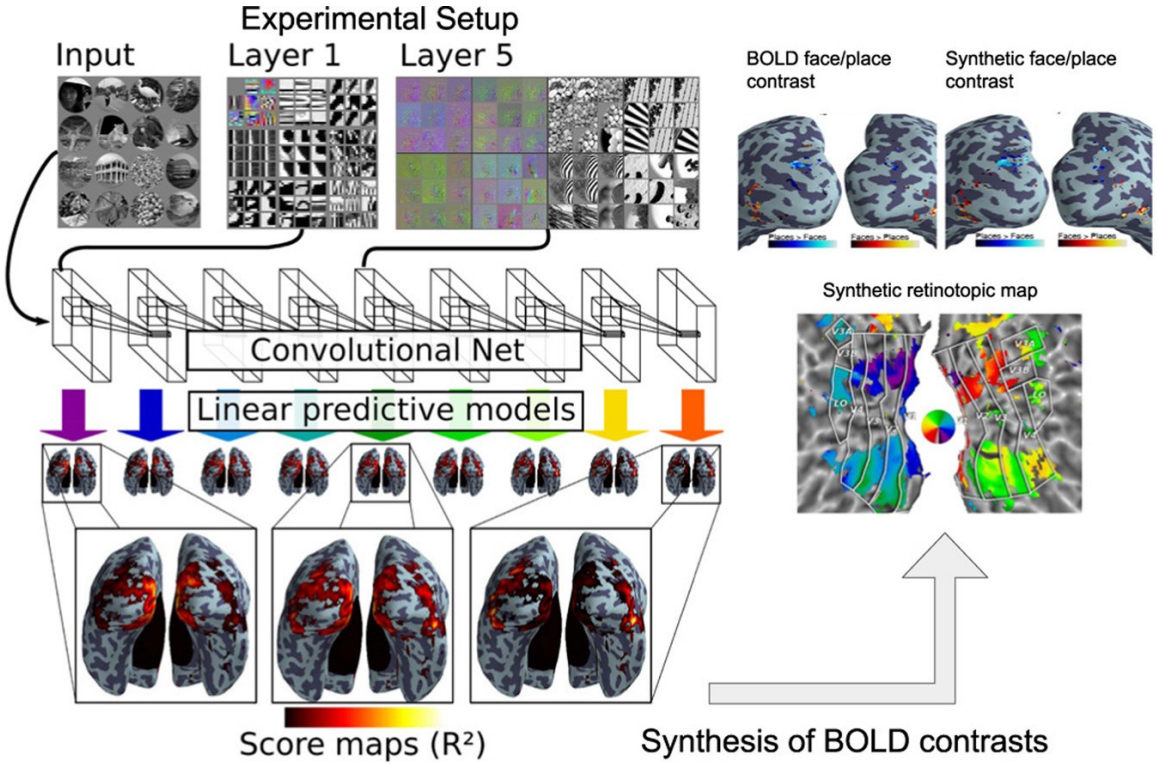
Neuromatch.io/neuroai-course/



# Deep learning as a plausible cognitive architecture

- **Empiricist motivation**
  Cognitive structure must be **learned from experience**, requiring mechanisms that can reorganize in response to environmental regularities. Architectures based on fixed symbols or innate rules cannot do this; learning-driven architectures can
- **Connectionist solution**
  Connectionism models cognition as emerging from **simple interacting units shaped by experience**, allowing perceptual and conceptual structure to form gradually
- **Deep learning as instantiation**
  Deep learning implement this connectionist architecture. In vision, hierarchical DNNs develop **low-dimensional, decodable representations** that align with meaningful variables (e.g., systematic correspondences between model layers and human visual areas in Eickenberg et al., 2017)
- **Inference under cognitive criteria**
  Under explicit cognitive-interpretation criteria (latent structure, decodability, functional use, stability), DL architectures **cannot be ruled out in principle** as candidates for modeling human cognitive functions

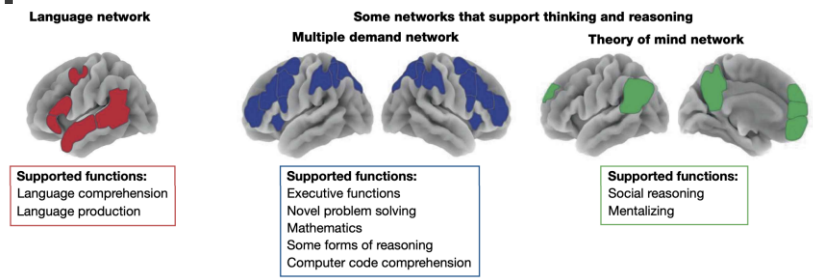
Eicklenberg et al., 2017. *NeuroImage*.




When the same natural images are shown to both humans (measured with fMRI) and a trained network, **early model layers best predict early visual areas (like V1/V2)** and **deeper layers best predict higher visual areas (like inferior temporal cortex)**.

**Open question:** Whether LLMs satisfy these criteria beyond perception and language remains an empirical, domain-specific question

# Is it only a matter of DL architecture?




Fedorenko et al., 2024. *Nature*.

**Large language mistake**

Cutting-edge research shows language is not the same as intelligence. The entire AI bubble is built on ignoring it.

by Benjamin Riley
Nov 25, 2025, 7:00 AM EST

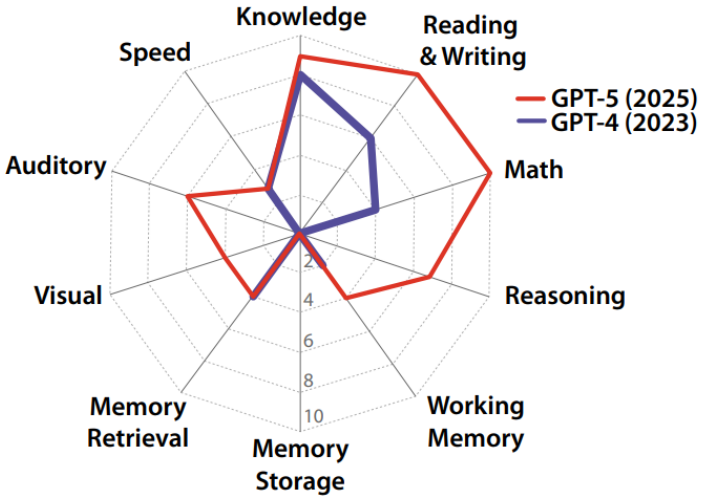
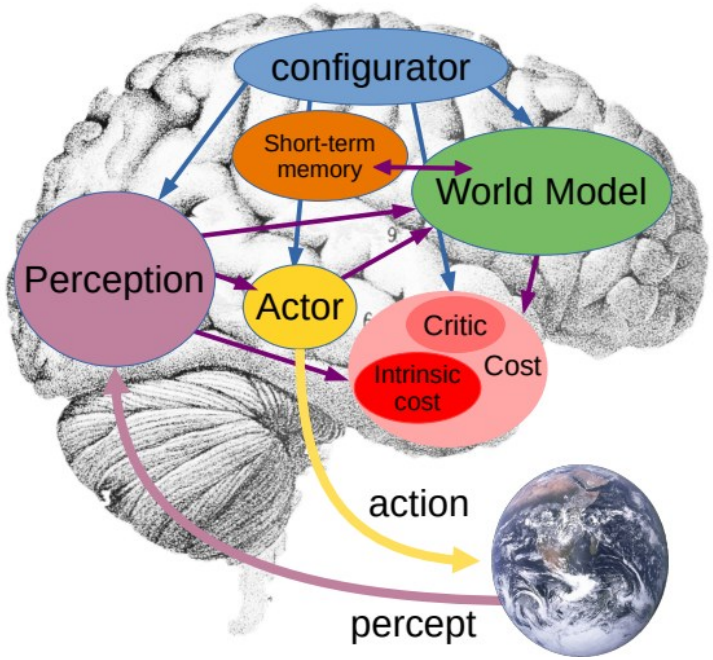



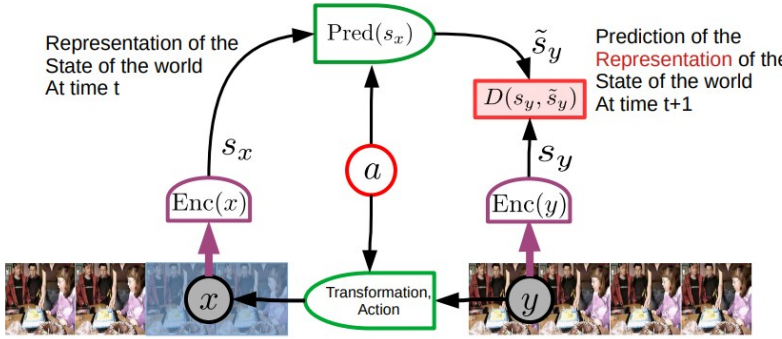

Figure 1: The capabilities of GPT-4 and GPT-5. Here GPT-5 answers questions in 'Auto' mode.

Hendrycks, Bengio et al., 2025. *arXiv*.

LLMs cannot perform online learning

Sutton on Dwarkesh Pattel podcast, September 2025




World models & Joint Embedding Predictive Architecture (JEPA),

LeCun 2022. *Open Review*.

# The lost art of deriving theoretical implications

## Closing statement

- **We will need to clarify what we mean by a cognitive function**
  Whether we define a function behaviorally, computationally, or representationally changes how we evaluate cognition in both humans and machines
- **Cognitive functions must be expressed in operationalizable terms to be testable**
  Once defined this way, we must acknowledge that machines performing computations may be legitimate cognitive models (and perhaps cognitive agents). Second, comes the need to define when a theoretical concept of cognitive function is transposable from human to machine, and from machine to human
- **This requires making our theoretical commitments explicit**
  Do we think the brain represents? Predicts? Is cognitively penetrable? Follows 4E principles?
  Can we expect the same from a machine?
  Our answers determine which model architectures count as plausible, and informative
- **Frameworks and paradigms provide mechanisms for how structure can emerge from inputs**
  They explain how systems learn concepts, abstractions, and hierarchical organization. They give us a set of assumptions and a set of constrains to navigate the space of claims
- **We somewhat lost the art of deriving the theoretical implications of our own assumptions and findings**
  Models and benchmarks improved, but the field paid less importance to anchoring them in coherent paradigms. Reclaiming theoretical rigor is essential for making meaningful cognitive claims about any form of intelligence