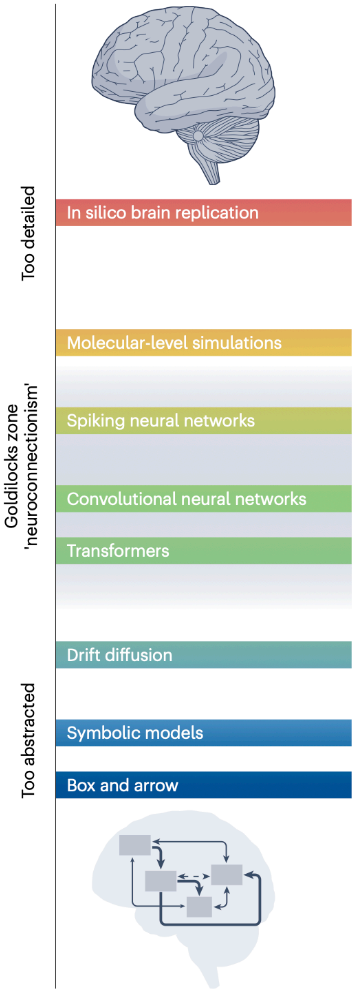
Jun 30, 2024 Neuroconnectionist Research Program

The task of cognitive computational neuroscience is to find **the right level**, with enough fidelity to biology to preserve the essential mechanisms, but abstract enough to discard details not required for cognitive function, which reproduces the trajectory from actively sensed input, through internal representations realized in neural processes, to complex goal-directed behaviors. **Abstraction is central to reveal what details matter and what approaches closer to the truth.**

Imagination may be a limitation, what we can research about the brain may be limited to what we can come up with, but natural mechanisms are not necessarily bound within these constraints: neural selectivity can often rely on more complex features that only imperfectly map onto human-interpretable categories. Inorder to give a complete picture of how cognition emerges, brain science needs interpretable computational models that **go beyond the limits of human-interpretable labels for neural activity**, that are applicable in naturalistic settings by being grounded in sensory data and that tie together multiple levels of explanation. This new approach is termed to be ‘neuroconnectionism’ — a cohesive large-scale research programme centered around ANNs as a computational language for expressing falsifiable theories and hypotheses about multi leveled brain computation.



Neuroconnectionism has already been successfully applied in a wide variety of neuroscientific settings, including vision, audition, semantics, language, reading, **decision-making**, **attention**, **memory**, game playing, **motor control** and the formation and coding principles of brain areas.

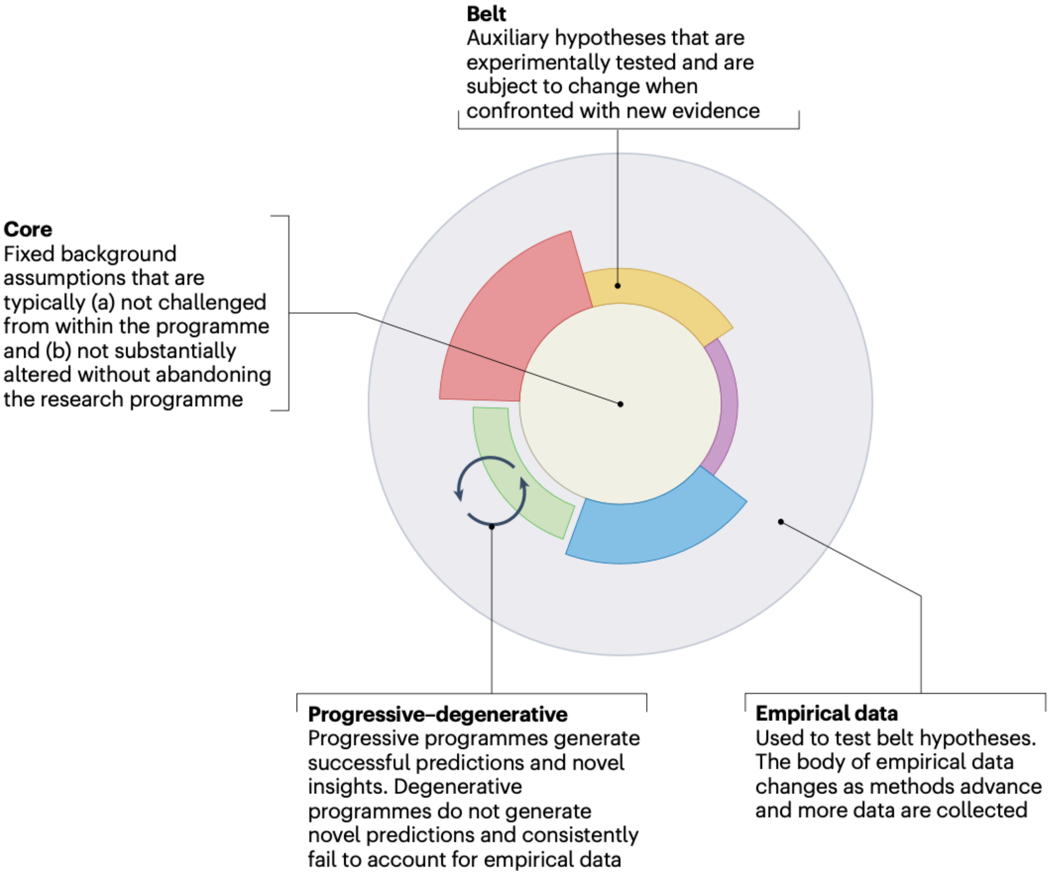
The search for inhabitable exoplanets means looking for planets orbiting at the ‘right’ distance from their stars to have liquid water. If they are too close, temperatures are too high and water evaporates. If they are too far, temperatures are too low and water freezes. The temperature has to be just right, as in the Goldilocks fairytale. Analogously, models that are too close to the biological brain fall outside the Goldilocks zone because they have too much biological detail and cannot be run or trained at scale to perform complex cognitive tasks from sensory grounded evidence. Models that are too abstract also fall outside the Goldilocks zone as they can neither be easily linked to biology, nor be grounded in sensory input. As unnecessary detail complicates understanding, models need to focus on incorporating the biological elements crucial for explaining brain computation at an appropriate level of abstraction.

Lakatosian Research Program

The neuroconnectionist research program is built around the Laktosian perspective of scientific philosophy, which differs from the mainstream idea that people used to believe what science is. It forms the core of belief in the center of the program and expands auxiliary hypotheses that can be tested to the side of it.

Popperian: theories are rejected when they are falsified in tests was dominant: - If T, then O - Not O, hence not T Where T is the theory and O is the observation	**Laktosian**: give what we found some confidence. It is about our belief of how the world works, don't reject it unless when we have to: - If T, and A1, and, …, and An, then O - Not O, hence not T, or not A1, or not, …, or not An where (A1, …, An) are auxiliary hypothesis Argues that science would not work like the Popperian idea **in practice** and could not work like that **in principle**. It is never a single hypothesis, but a whole collection of hypotheses that generates predictions, any one of which might be at fault if the prediction is not vindicated. The (not O) may not come from the wrong T, but the wrong A1 to An that surrounds T. 1. Belief at the outer circle describes localized observable facts and the ones in the center describe some sense of generalized belief. 2. A theory is rejected not as the result of a direct conflict with the evidence, but **because the attempt to preserve the core principles becomes so cumbersome** that they cease to form a productive working hypothesis for continued testing and the discovery of new insights. 3. When such conditions happens, the research paradigm need to be changed and **a complete overhaul of theories and the language used to describe the world would need to be changed** (Kuhnian scientific paradigm shift).

In astronomy, deviations in planetary trajectories from the smooth ellipses predicted by Newtonian mechanics were observed. Instead of rejecting Newtonian laws owing to these challenging empirical data, scientists assumed the correctness of the laws and tested auxiliary hypotheses (such as the presence of an unseen planet) that might explain the orbital deviations. Hence, a belt claim was falsified (the number of planets in the solar system) but the core was not abandoned (Newtonian mechanics). Then in the twentieth century, evidence accumulated against Newtonian celestial mechanics that could not be solved assuming the correctness of the laws, which led to its rejection and the development of general relativity, a novel progressive core that changed the way the universe is thought about and led to great discoveries.



Belt
Auxiliary hypotheses that are experimentally tested and are subject to change when confronted with new evidence

Core
Fixed background assumptions that are typically (a) not challenged from within the programme and (b) not substantially altered without abandoning the research programme

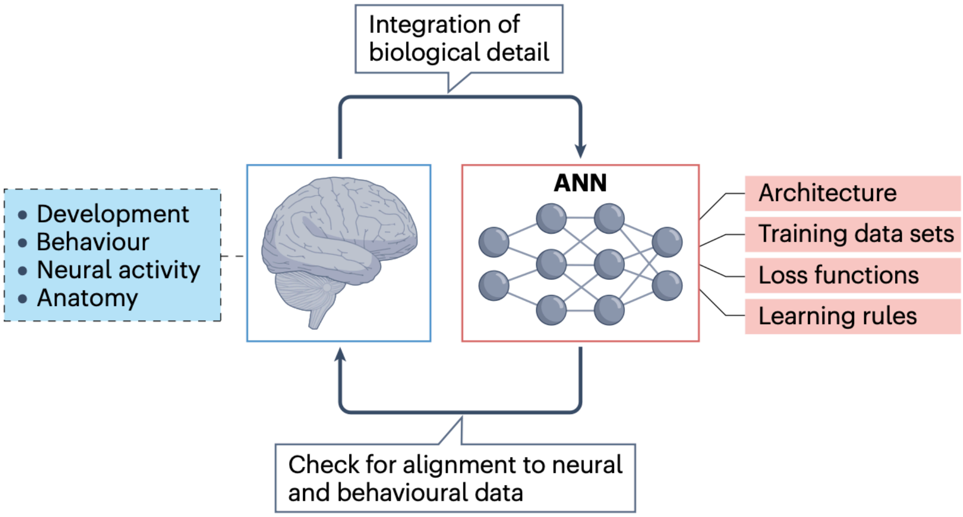
Progressive–degenerative
Progressive programmes generate successful predictions and novel insights. Degenerative programmes do not generate novel predictions and consistently fail to account for empirical data

Empirical data
Used to test belt hypotheses. The body of empirical data changes as methods advance and more data are collected

Core of Neuroconnectionist

1. Accept the fact that the brain is complex and brain science requires complex, distributed and iterative models to reveal the true mechanism and theory of how the brain operates. Analytical solutions would not exist, complex systems need complex tools to map.
2. ANNs offer a highly suitable computational language: sufficiently abstract to be computationally tractable and reproduce cognitive functions, while still being close enough to biology to relate to, implement and test neuroscientific hypotheses.

Belt of Neuroconnectionist



1. Architecture + Data + Objectives + Learning Rules
   a. Random reservoirs, convolutional layers, inductive biases (memory candidate)
   b. Supervised (classification and scene captioning), unsupervised (contrastive learning, predictive coding, image generation, temporal stability, and energy efficiency), and behavioral reward
   c. Backward propagation, Hebbian learning, predictive coding, self-organizing maps.
2. Behavior + Neuronal + In Silico Physiology + Developmental Agreement
   a. Representational Similarity Analysis (RSA) looks at the populational representational geometric similarity.
   e. Use linear combination with activations to predict neuronal activity.
3. Math & Neuroscience may be somewhat connected, providing mathematical theoretical insights to how the brain might be working.
   a. As ANNs are heavily overparameterized and learn non-convex loss functions, precise mathematical tools are required to better understand the underlying computations and learning dynamics (deep mathematics). Insights from deep mathematics (double descent, neural tangent kernels) are of great importance for understanding complex neural processes, as the brain, too, is highly overparameterized.

Goldilocks Zone & Problems Makes Developments

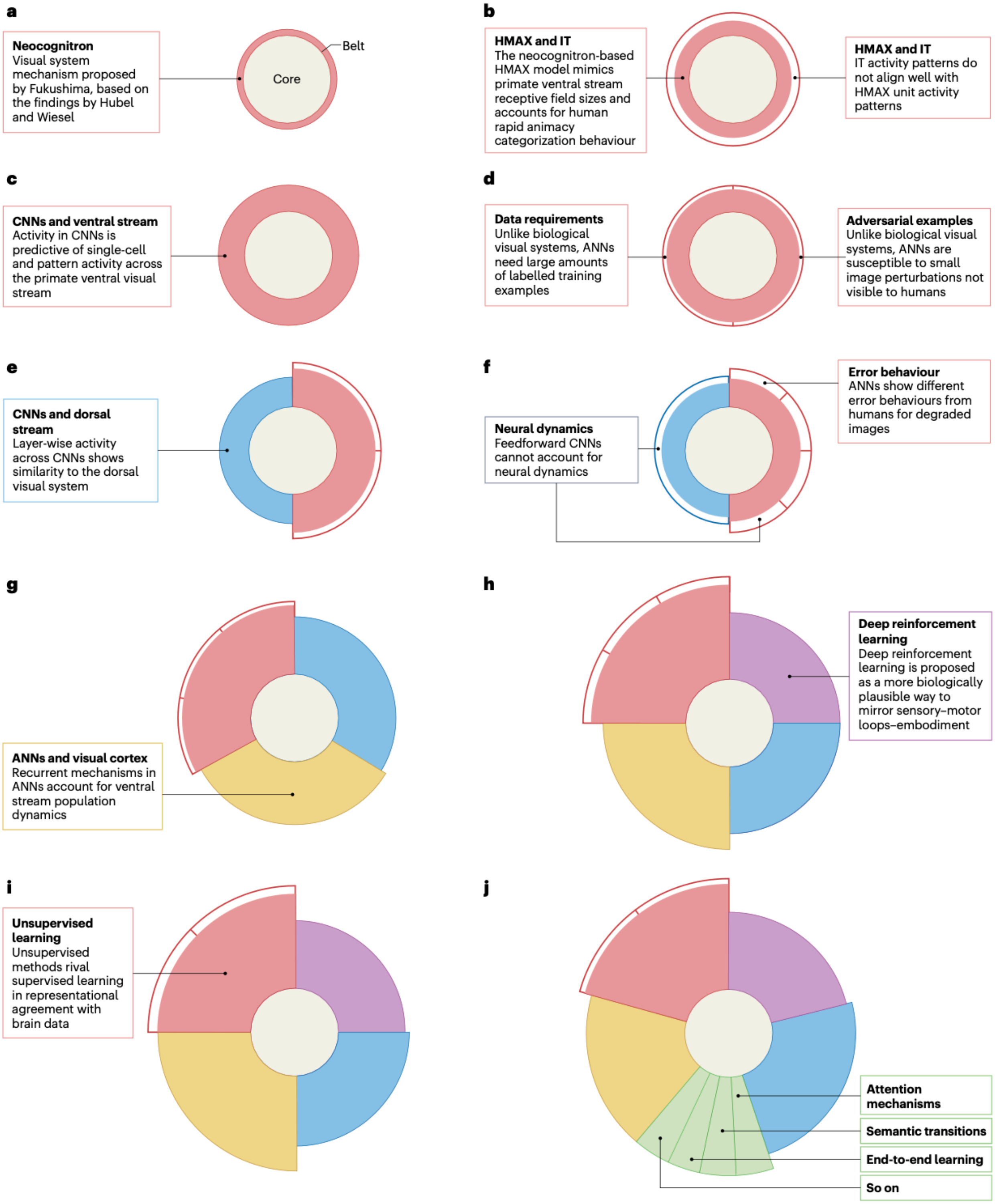
ANNs live in the Goldilocks zone of biological abstraction, **striking the required balance between biological realism and algorithmic clarity, providing a level of abstraction much closer to biology but abstract enough to model behavior**. They can be trained to perform high-level cognitive tasks, while they simultaneously exhibit biological links in terms of their computational structure and in terms of predicting neural data across various levels — from firing rates of single cells, to population codes and on to behavior.

Individual elements of the belt are important, but a more central aim, when taking a Lakatosian perspective, is an **evaluation of longitudinal developments** (both theoretical and empirical), which determine whether a research programme is progressive or degenerative. New hypotheses can be derived and existing hypotheses can be corroborated, altered and rejected so that the belt of a research programme is subject to change.

An individual belt hypothesis that is rejected **does not refute the core assumptions upon which a research programme is built, but rather provides an important datapoint for future developments**. According to this Lakatosian view, the overarching question becomes: How does the neuroconnectionism research programme fare in terms of productivity, discussing whether neuroconnectionism generates new insights, and how well it addresses existing challenges.

Historical development in VNL that have been expanding the belt:

1. **(A,B) Neocognitron** was derived from seminal findings about simple and complex visual system cells by Hubel and Wiesel. It learned and recognized increasingly abstract visual patterns through mechanisms that were similar to convolutions. **HMAX** is a more powerful model and was shown to match well to human psychophysical data on animacy detection well but did not align well with broad activity patterns observed in IT, providing disconfirmatory evidence and weakening the belt item.
2. **(C,D,E,F) CNNs** layer activities match neural activity patterns along the primate ventral visual stream.
    a. First time that a single image-computable and functional object recognition network was able to match activity patterns across the ventral visual system.
    b. They have susceptibility to adversarial attacks and the amounts of labeled training data they required, which were shown to exhibit several important differences with biological vision.
    c. Have similar layer activities to the dorsal visual stream.
    d. Error behavior during image alterations diverges between humans and CNNs.
    e. Feedforward CNNs embodied too simple mechanisms to cover neural dynamic observations beyond coarse rate coding.
3. **(G)** Dynamic transformations during visual processing can be captured if **recurrence** is added to ANNs.
4. **(H,I)** Activity across the dorsal visual stream during game playing matches activity in **deep reinforcement learning** networks, which implement a sensory–motor loop for the same game playing tasks. **Unsupervised learning** can rival supervised learning in representational agreement with brain data, which solved the challenge that too many labeled examples were needed for training.
5. **(J)** Future directions: Attention mechanisms, semantic objectives and end-to-end learning in which networks are trained directly to match neural activity are recent developments in ANNs.



Neocognitron
Visual system mechanism proposed by Fukushima, based on the findings by Hubel and Wiesel

HMAX and IT
The neocognitron-based HMAX model mimics primate ventral stream receptive field sizes and accounts for human rapid animacy categorization behaviour

HMAX and IT
IT activity patterns do not align well with HMAX unit activity patterns

CNNs and ventral stream
Activity in CNNs is predictive of single-cell and pattern activity across the primate ventral visual stream

Data requirements
Unlike biological visual systems, ANNs need large amounts of labelled training examples

Adversarial examples
Unlike biological visual systems, ANNs are susceptible to small image perturbations not visible to humans

CNNs and dorsal stream
Layer-wise activity across CNNs shows similarity to the dorsal visual system

Error behaviour
ANNs show different error behaviours from humans for degraded images

Neural dynamics
Feedforward CNNs cannot account for neural dynamics

ANNs and visual cortex
Recurrent mechanisms in ANNs account for ventral stream population dynamics

Deep reinforcement learning
Deep reinforcement learning is proposed as a more biologically plausible way to mirror sensory–motor loops–embodiment

Unsupervised learning
Unsupervised methods rival supervised learning in representational agreement with brain data

Attention mechanisms

Semantic transitions

End-to-end learning

So on