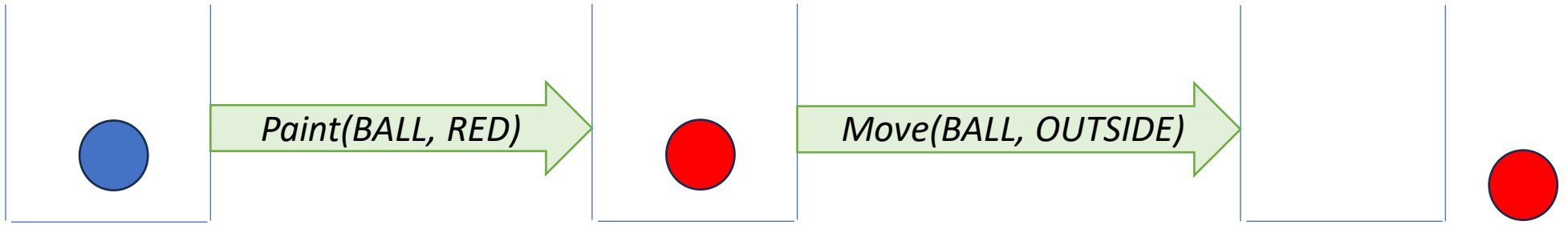
Epistemology of AI

HSL622/ELL457

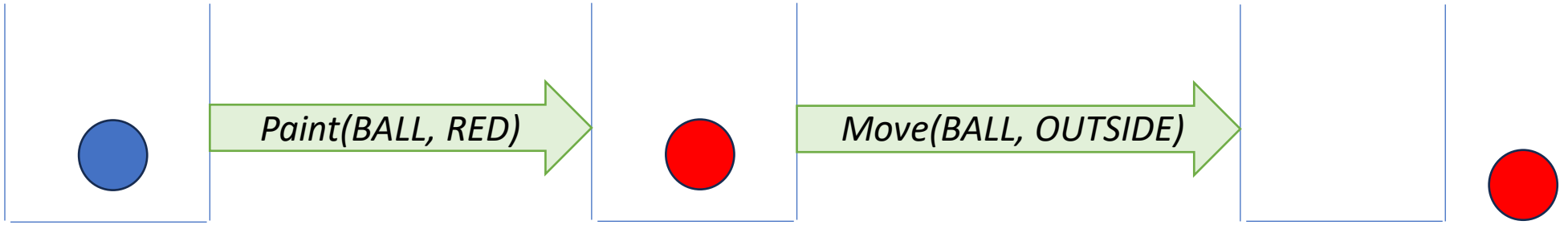
Key questions

- What kinds of things can an AI system come to know, and under what conditions?
- How does this depend on the nature/components of the AI system (e.g., symbolic vs. connectionist; domain-specific vs. domain-general)?
- Are there specific epistemological challenges we can identify for particular types of AI systems?
- Where are the *gaps* between what AI systems can know and what humans can know?
- What kind of cognitive architecture can account for the epistemic capacities of humans? Can it be fully characterised computationally?

Classical AI: *The Frame Problem*



Classical AI: *The Frame Problem*



- How can such an AI system know that the *Move()* action doesn't affect the colour of the object being moved? Or that *Paint()* doesn't affect the location?
- If these have to be included as part of the definition of *Move(), Paint()*, etc. (known as frame axioms), then isn't there an unbounded number of such axioms or non-effects of every action?

[Example adapted from Murray Shanahan, 2016]

More general epistemological frame problem

- We (or cognitive agents generally) may have many beliefs about the world; more generally, many intentional states
- If a given action is carried out in the world, or a given piece of information received by the agent, *which* beliefs should it update? [Dennett 1978, Fodor 1983]
- This is a problem of relevance: given a large (potentially unbounded) number of intentional states, and a large (potentially unbounded) number of dynamical events that can occur in the world, how to determine the mapping of *which* states need to be updated in light of *which* events, given that the agent wants the states (or their intentionality) to retain *faithfulness* to the world?

Isotropic nature of relevance

Belief: Margarine is a good substitute for butter on my breakfast toast

Event in the world: Riots in Borneo due to large-scale deforestation

Is this event relevant to the status of this belief?

Isotropic nature of relevance

Belief: Margarine is a good substitute for butter on my breakfast toast

Event in the world: Riots in Borneo due to large-scale deforestation

Is this event relevant to the status of this belief?

It could be, if margarine is known to contain palm oil sourced from Borneo.

Points to computational intractability of the determination of relevance – need to check all stored representations of information?

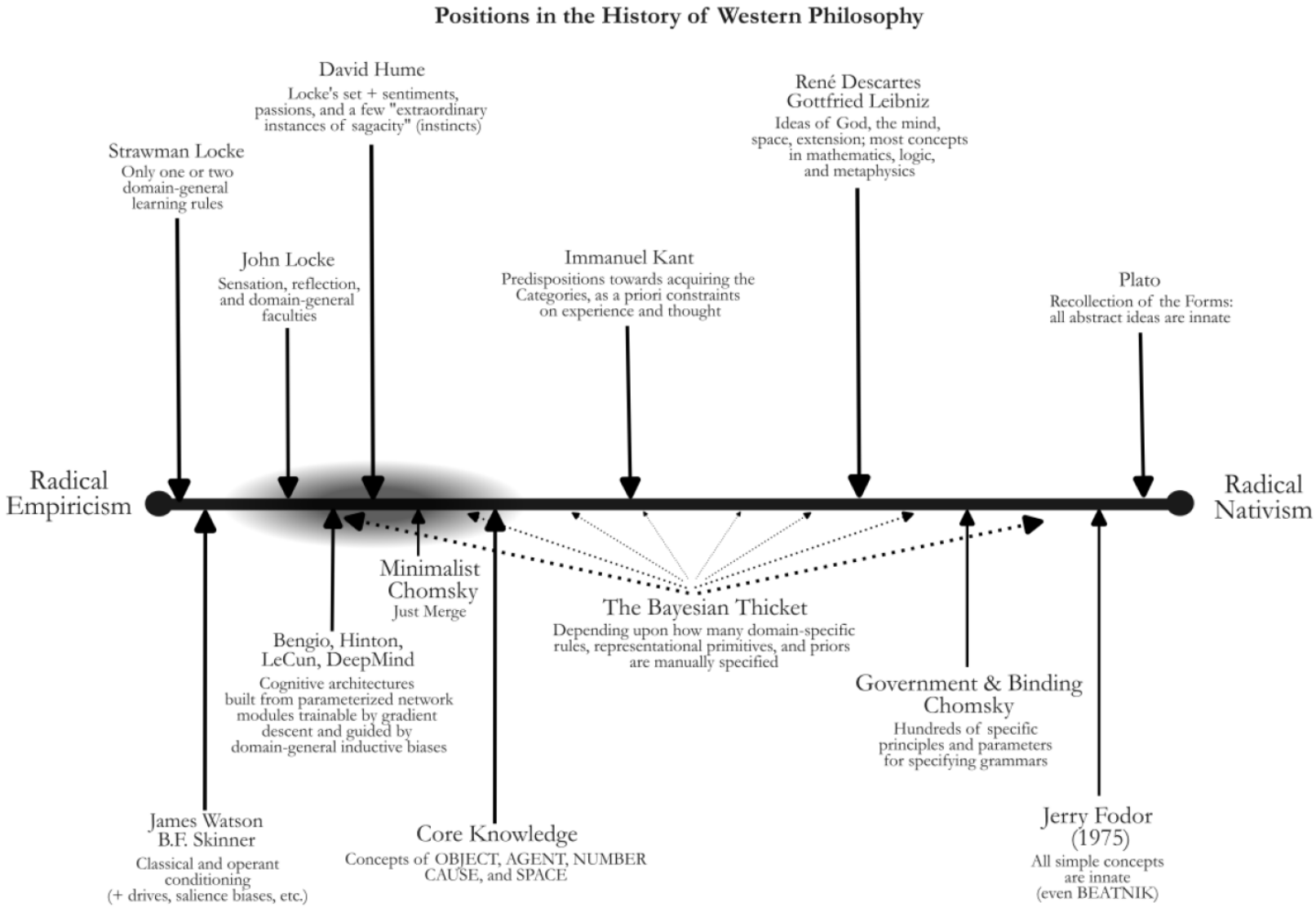
[Example adapted from Nicholas Shea 2024]

Do connectionist models solve the (broader) frame problem?

- Hardly discussed in AI today; perhaps neural network / deep learning models are effectively able to deal with it, though no one entirely understands how
- It may come at the cost of a *very* large number of weights/parameters, which can be seen as encoding association strengths between different representations [GPT-4 is rumoured to have 1.76 *trillion* parameters]; all these parameters along with a large context window might provide a way to effectively infer/obtain relevant context from the input itself
- Such intensive computational complexity can be seen as reflecting what Nicholas Shea [2024] calls the *If-Then Problem*, or C. R. Gallistel the *Infinitude of the Possible* [Gallistel and King 2010]: perhaps just a variant of the frame problem?

Design of cognitive architectures

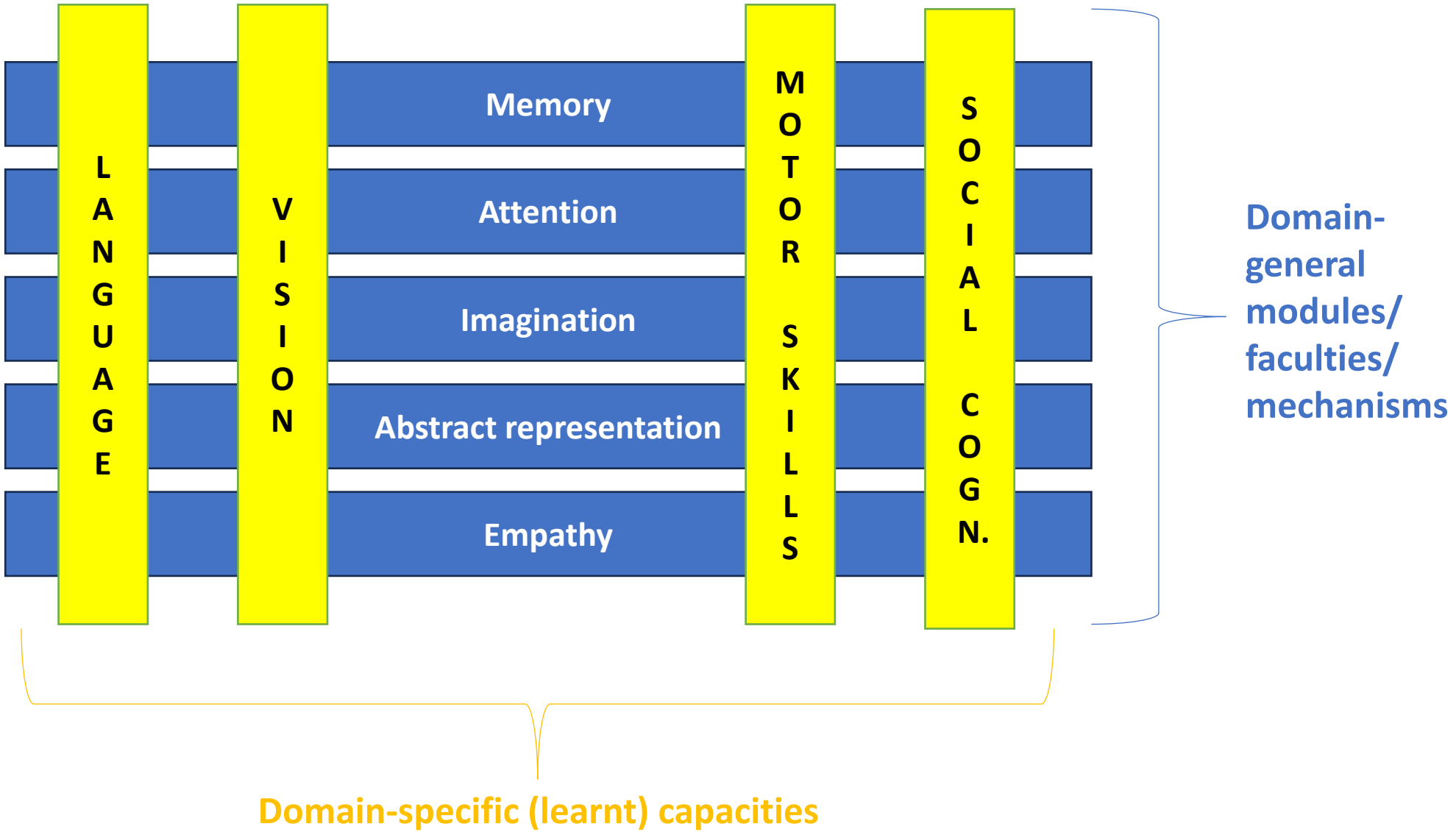
- A framework consisting of components/modules/mechanisms (typically computational) which can serve as a basis for *implementing* or *realising* or *simulating* various cognitive capacities
- The evolution of the use of different computational models or mechanisms in AI and Cognitive Science has some important parallels with the more general nativism vs. empiricism debate in epistemology and philosophy of mind [Buckner 2023]
- In particular, symbolism and domain-specificity have tended to reflect more nativist choices; connectionism (especially deep learning) and domain-generality have tended to reflect more empiricist choices



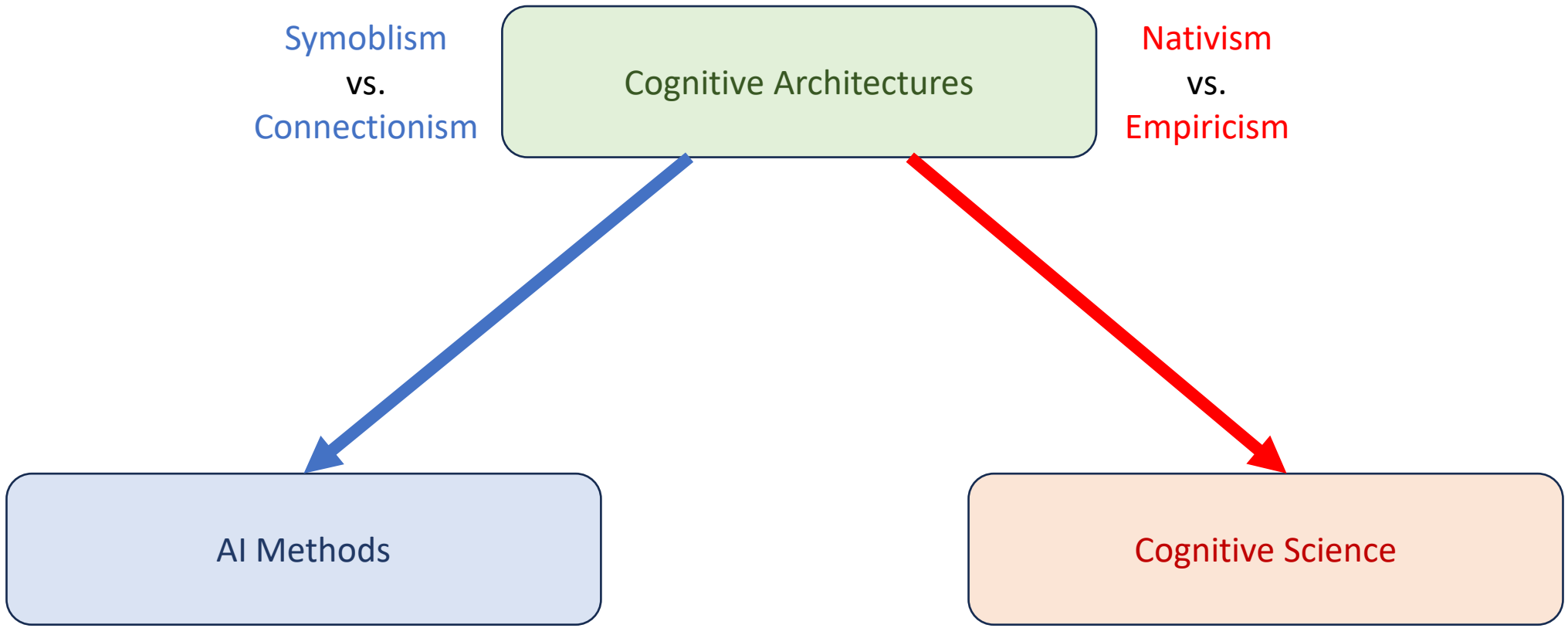
[Buckner 2023]

(Radical) Nativism	Moderate/Origin Empiricism [Buckner 2023]	(Radical) Empiricism
Domain-specific building blocks	Domain-general modular architecture (DoGMA)	Very general/universal learning mechanisms
Innate modules/faculties/mechanisms (e.g., language faculty)	Specific faculties to be found empirically	No modules per se
Limited model-based learning: strong starting model with innate concepts	Extensive model-based learning, no specific innate concepts/representations: may mirror deep learning	All learning, *tabula rasa* (?): no or very minimal starting 'model' of world

DoGMA

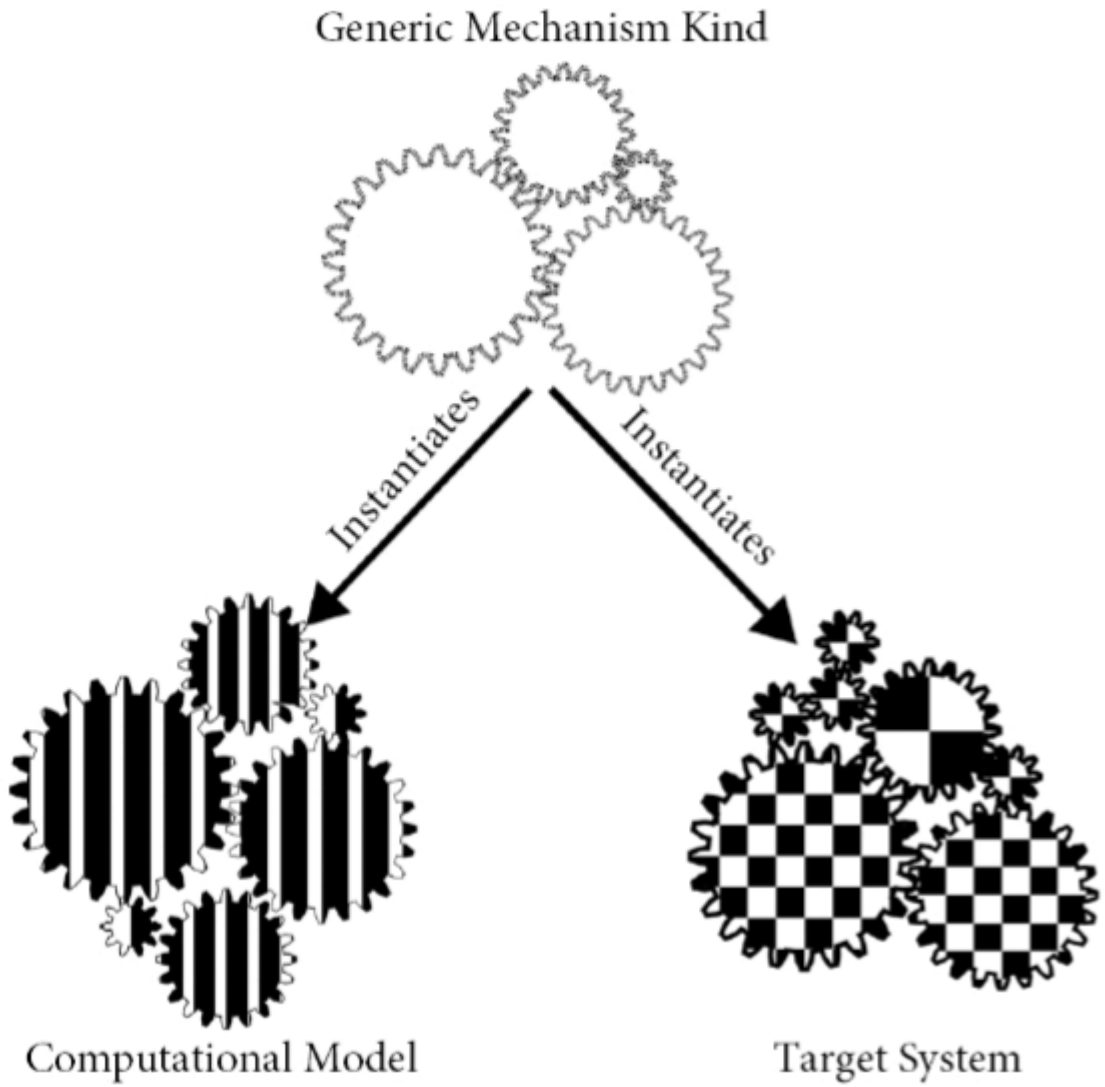


[Schematic representation of Buckner 2023]

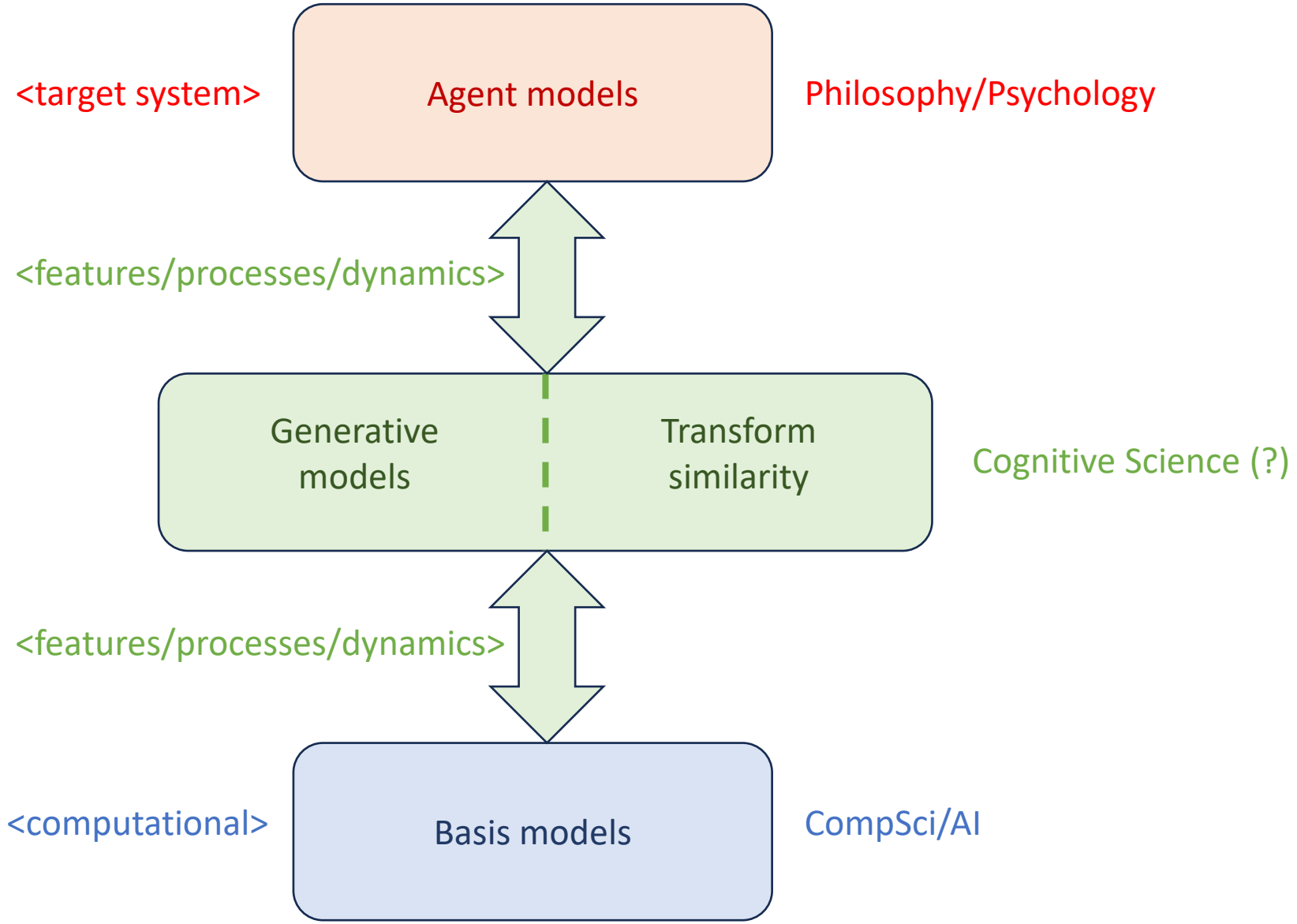

Symoblism
vs.
Connectionism
Cognitive Architectures
Nativism
vs.
Empiricism
AI Methods
Cognitive Science

Model-Mind correspondence?

- One kind of correspondence is just at the level of *behaviour* or *performance* – typically the goal in AI
- Arguably, this is *mimicry* and not necessarily mechanistic emulation or explanation [Stinson 2020; Jaeger 2023]
- A stronger criterion for cognitive science might be correspondence at the level of *generic mechanisms*, or shared membership in a kind of abstract mechanistic structure or class [Stinson 2020]
- But maybe this is too demanding; maybe enough if certain features/processes/dynamics can be *mapped* from model to mind? [Miracchi 2019; Cao and Yamins 2021]



[Reproduced from Cameron Buckner 2023]



[Schematic representation of Miracchi 2019; Cao and Yamins 2021]

References

- Shanahan, M. The Frame Problem. *The Stanford Encyclopedia of Philosophy* (2016).
- Dennett, D. *Brainstorms*. MIT (1978).
- Fodor, J.A. *The Modularity of Mind*. MIT (1983).
- Shea, N. *Concepts at the Interface*. Oxford (2024).
- Gallistel, C. R. and King, A. P. *Memory and the Computational Brain: Why Cognitive Science Will Transform Neuroscience*. Wiley-Blackwell (2010).
- Buckner, C. J. *From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence*. Oxford (2023).
- Stinson, C. From implausible artificial neurons to idealized cognitive models: Rebooting philosophy of artificial intelligence. *Philosophy of Science* **87**(4):590–611 (2020).
- Jaeger, J. Artificial intelligence is algorithmic mimicry: why artificial "agents" are not (and won't be) proper agents. *arXiv*:2307.07515 (2023).
- Miracchi, L. A competence framework for artificial intelligence research. *Philosophical Psychology* **32**(5):588–633 (2019).
- Cao, R. and Yamins, D. Explanatory models in neuroscience: Part 1 -- taking mechanistic abstraction seriously. *arXiv*:2104.01490 (2021).