Turing, Searle, Lovelace: Three Levels of Mental Attribution in AI

BAME Workshop – C-tests

2026 - June, 10

Preliminaries

- Sits at the intersection of philosophy, theoretical computer science and AI

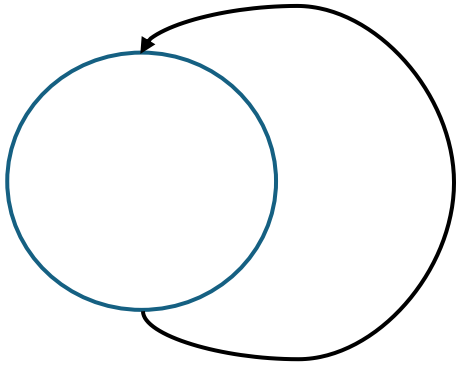
- Will not exclusively address phenomenal experience (too speculative in AI), but (artificial or “quasi”-)mental states and their attribution

Outline

- Pre-digital computers approaches to mind
- Testing the minds of machines (1950-2000)
  - Turing: behaviour
  - Searle: architecture
  - Lovelace: explanation
- Testing LLMs
  - Behavioral indistinguishability?
  - The impossibility of a static “rulebook”
  - LLMs in the Chinese Room & implications
  - LLMs & Lovelace
- Simulation vs Instantiation / Implementation
  - The “Abstraction Fallacy”
  - Physical Computationalism
  - Reconciliation
  - Behaviour / Architecture / Explanation revised
- Optimistic conclusion!

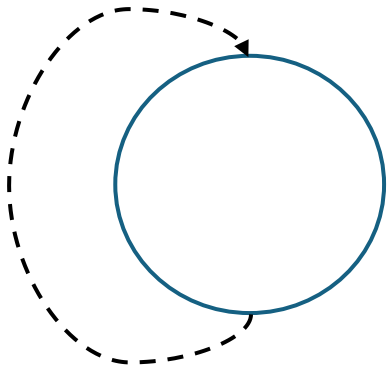
Descartes

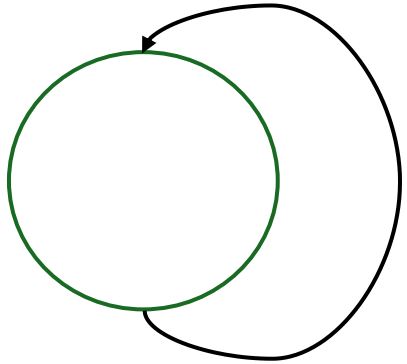
- “*Je pense donc je suis*”
- His approach to other minds:
  - Animals are not conscious
  - Machines can never be conscious:
    - Glitch: even if they perform convincingly, they will at times behave anomalously, revealing their artificial nature
    - “*Moral impossibility*”



Leibniz

- *“Calculemus!”*
- His approach to other minds:
  - Leibniz Mill: phenomenal experience is not directly visible in the mechanisms, no matter how far we zoom in





17th century

- Internal perspective
  - No proof needed because it is self-evidential through direct access
- External perspective
  - Phenomenal experience is private, not observable externally

Together, they raise serious concerns:

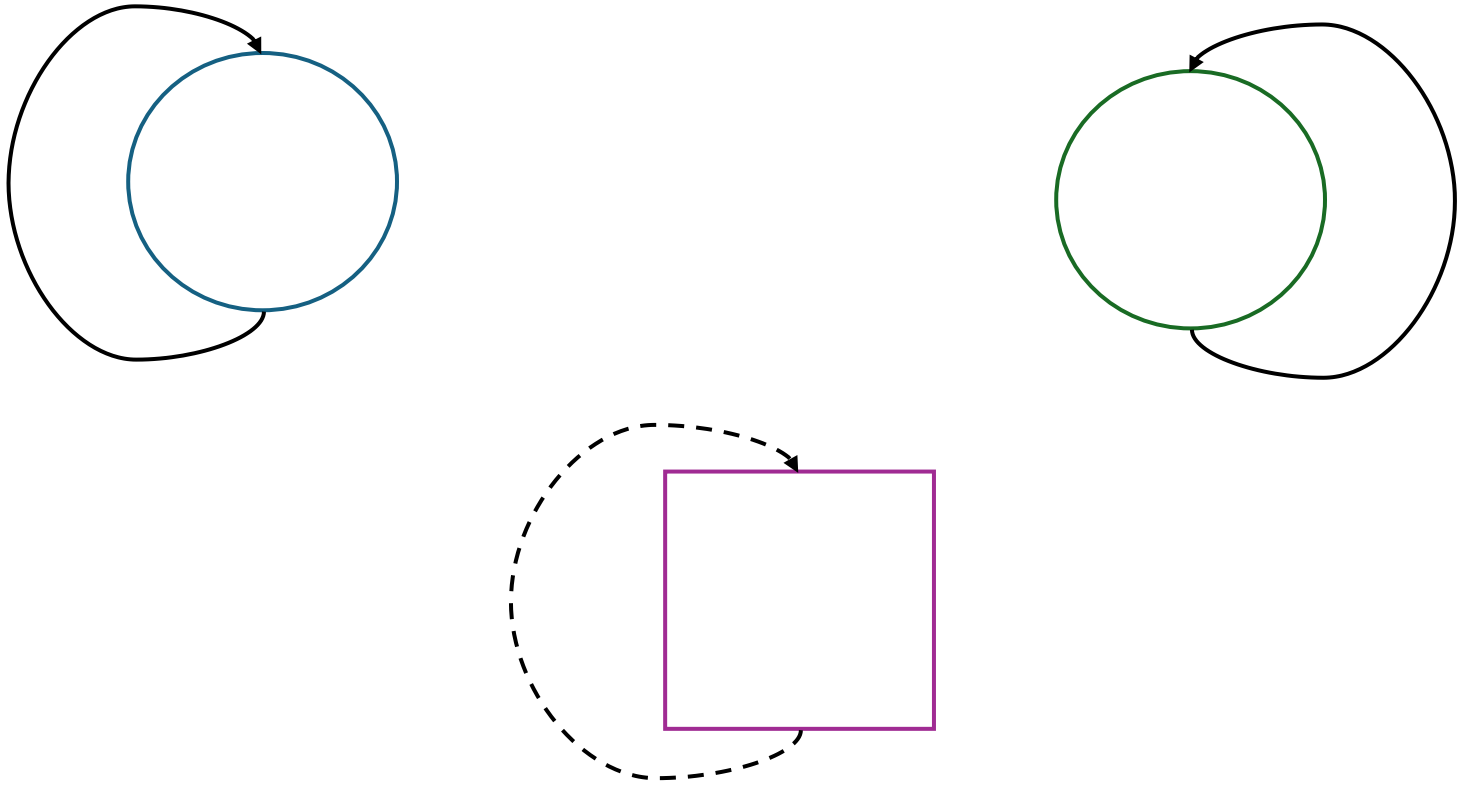
- Other minds problem
- Solipsism
- “where is the mind located”?

18th-mid-20th: birth of mind studies

- Birth of numerous disciplines to study minds and their abilities: psychology, psychiatry, psychoanalysis, psychism, etc…
- Problem: supernatural leanings, charlatanism, lack of methodological rigor, etc…
- “Mind studies” get a bad reputation unless they adhere to strict empirical canons

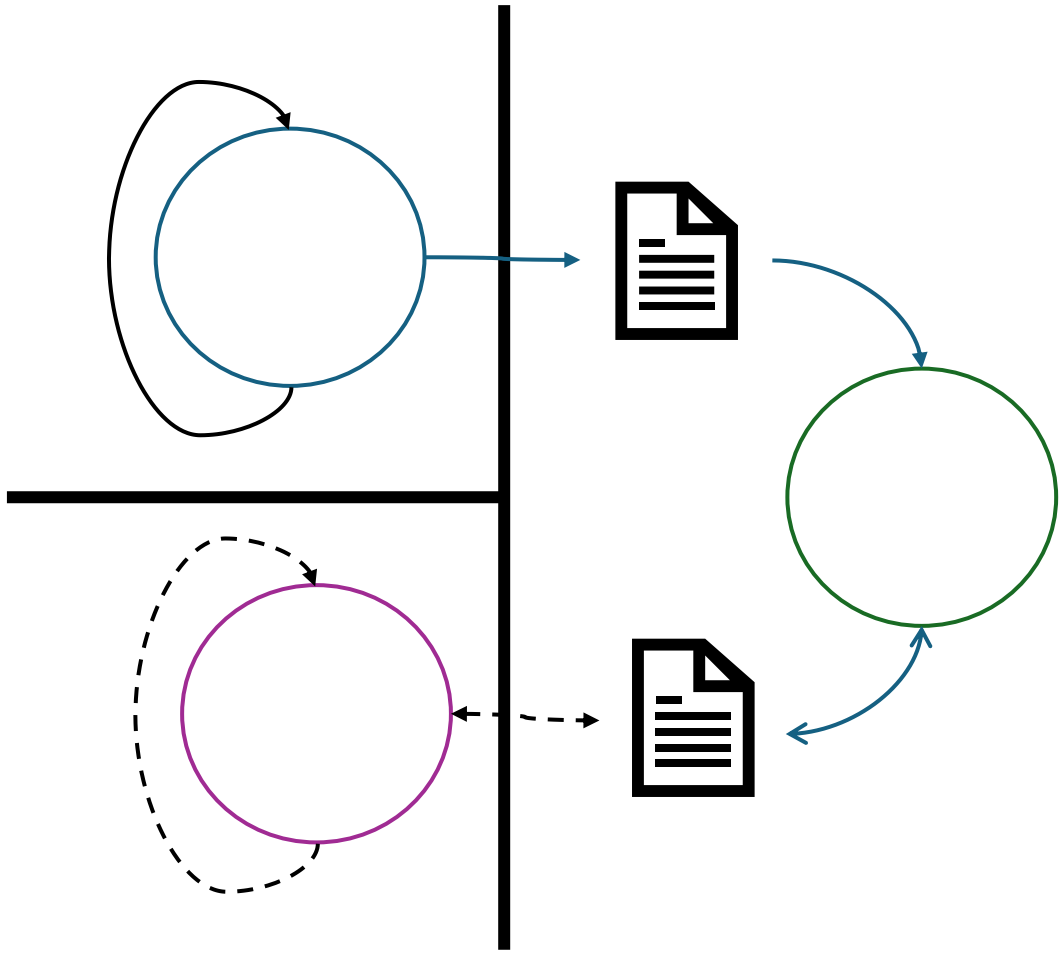
⇨Pragmatism, materialism and behaviourism explains away the problem of other minds

Mid-20th: digital computer enters the stage



Turing - Behaviour

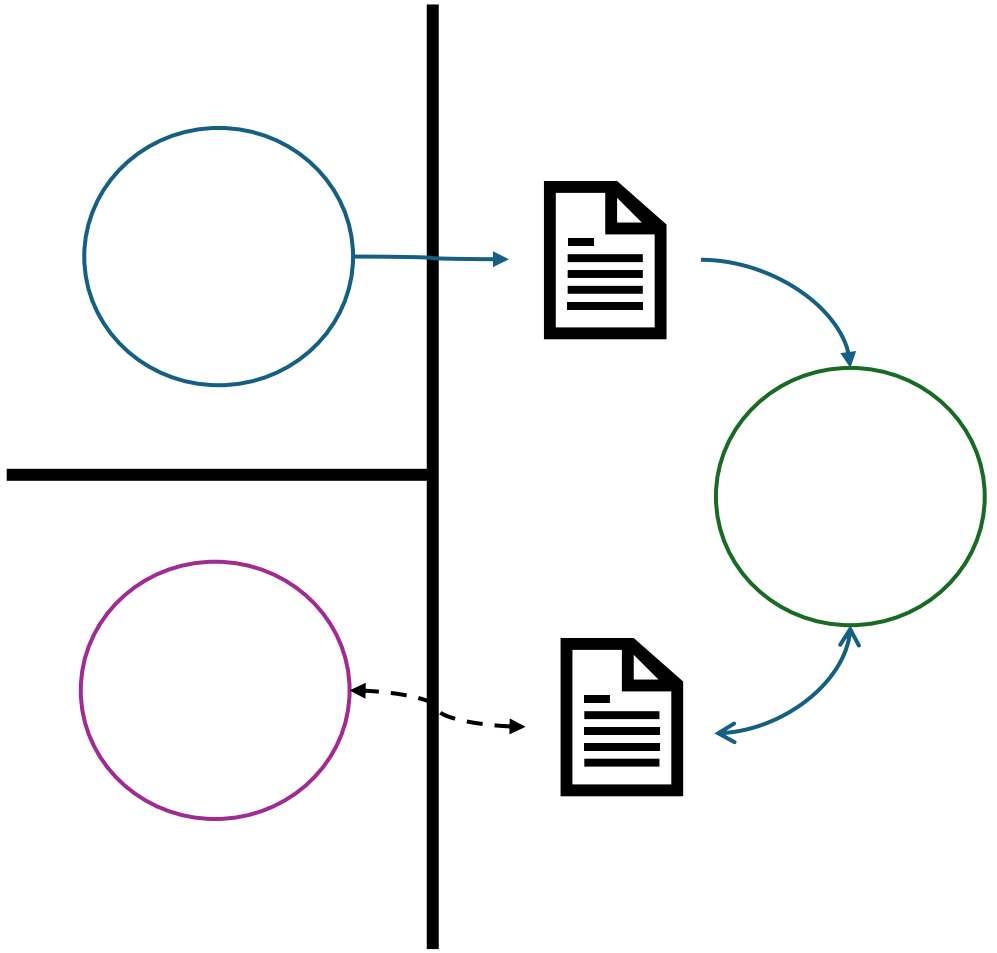
- Context: pre-AI
- Don’t ask if machines “think”, observe *what they do*
- Imitation game
  - Task: take turns in a typed conversation, pass for *X*



Turing - Behaviour

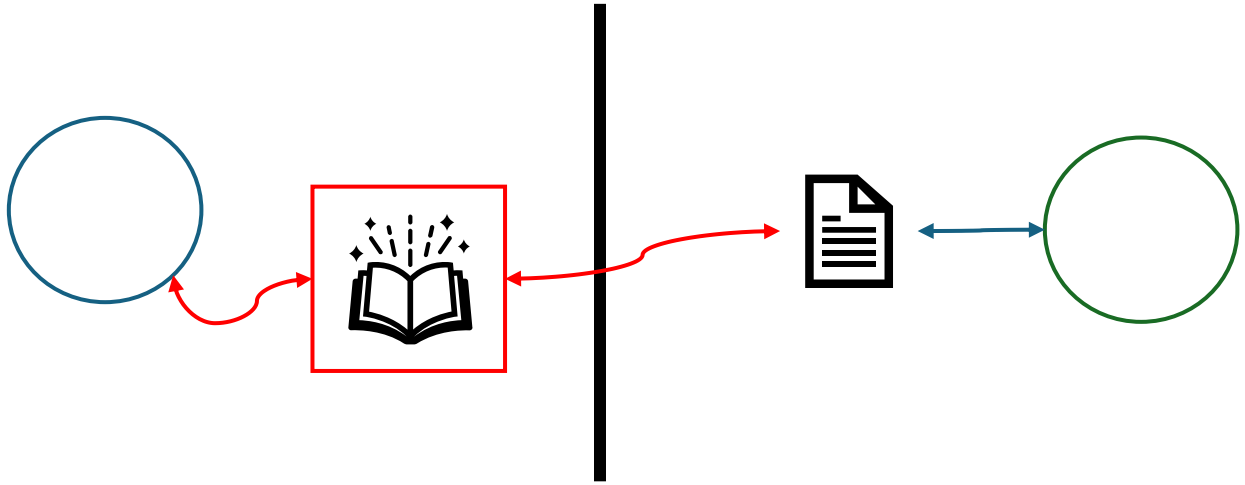
Underlying assumptions:

- The imitation game is substrate-flexible
- We don’t need to solve consciousness before we solve TT
- A digital computer can be programmed to type anything a human can



Searle - Architecture

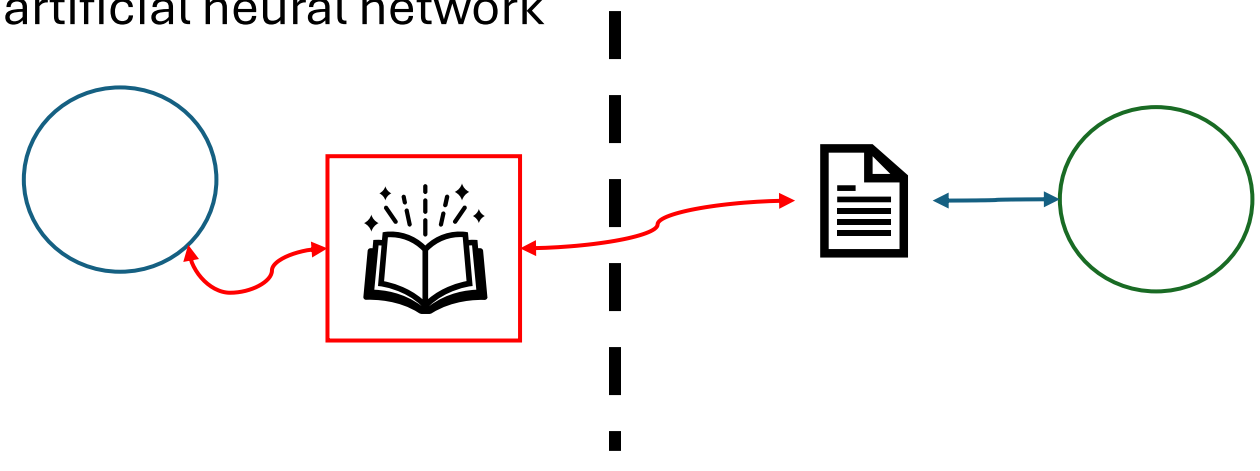
- Context: AI boom, cognitive turn, functionalism
- Machines may type, but they lack mental states
- Pb: Behavioural tests lead to false positives



Searle - Architecture

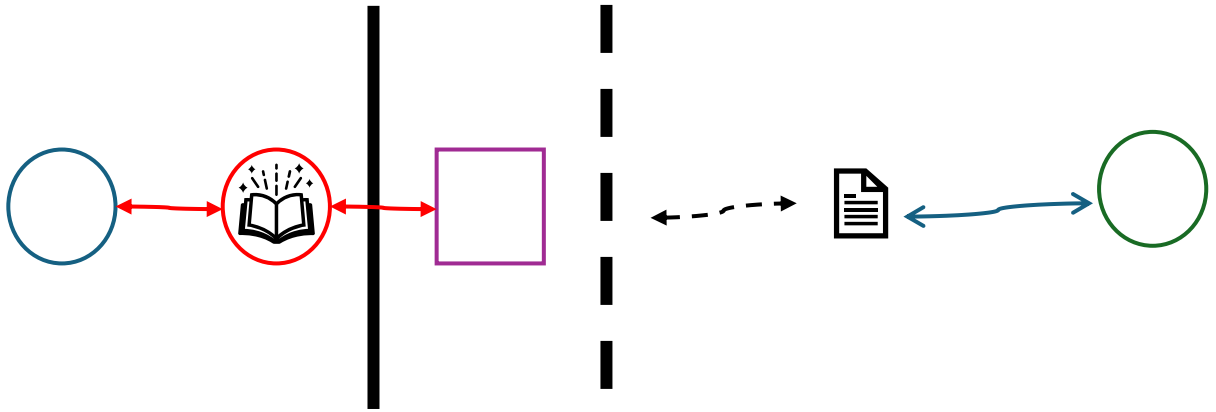
To evaluate mental states, one also needs to look *inside the room*

- How humans process language = semantics (*understanding*)
- How computers process language = syntax (*meaningless symbol manipulation*)
- Mental states are produced by the “causal mechanisms of a biological brain”
- Against functionalism: **digital computers *simulate***
  - Even by internalising the rules
  - Even with a robotic body
  - Even with an artificial neural network



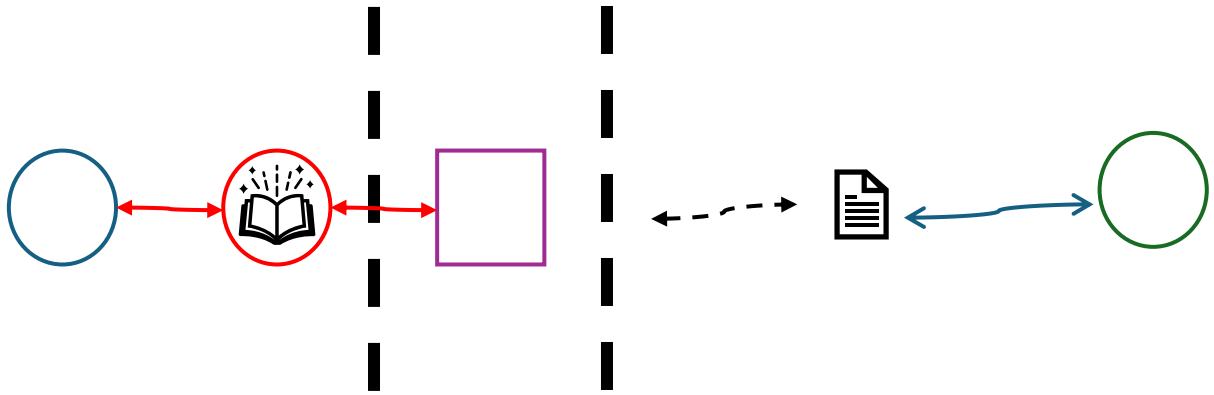
Lovelace - Explanation

- Context: 50 years after Turing, chatbots are still outputting answers following “Searle’s infamous rulebook”. (Bringsjord, Bello, Ferrucci)
- TT evaluators are not fooled by AI systems, but by their *designers*, the rulebook-keepers.



Lovelace - Explanation

- Intuition: is the behaviour caused by the architecture, or by the architect?
- Difficulty: slippery notions (causality, free will, …)
- Test for behaviour, architecture and ask whether the evaluator (or AI designer) can explain the AI's outputs by knowing how "the rulebook" was built



Do LLMs pass TT?

Yes:

- AI generates valid and relevant text.
- AI surpass human's performance at a number of linguistic tasks, including poetry, legalese etc…
- AI can impersonate a role: it is capable of fabricating a backstory and playing a human character.
- AI generated-text elicits emotional, intellectual and empathic responses from human evaluators (ignorant of the AI's nature).

Do LLMs pass TT?

No:

- AI is vulnerable to various techniques such as prompt-injections, where it processes text in a way that reveals their artificial nature.
- AI's context sensitivity and length, as well as memory across chats remain characteristic limitations.
- LLMs' self-reference is brittle: at times referring to themselves as AI, and at others using sentences such as "us, humans".
- Hallucinations, confabulations, sycophancy: their conversation flow is measurably different than a "healthy" human's.

Verdict

- Despite all these limitations of popular chatbot apps, if the goal is simply to fool a naïve human evaluator, LLMs can be tailored for that purpose in a convincing way.
- Under modest requirements, AIs pass the Turing test: they show us that computer programs can read and write meaningful sentences.
- This strongly suggests that chatting in natural language is a substrate-flexible task, which can be performed by digital computers.

The impossibility of a static “rulebook”

- The set of possible human conversations is:
  - Open-ended, living, difficult to define
  - Evocative, ambiguous, context sensitive
  - Combinatorially infinite
  - **Computationally intractable / too underspecified for exhaustive enumeration**
- Lookup tables
  - Matches each input with a predefined output stored in memory
  - Pb: not enough human time and physical space to write such an algorithm
  - **FAILS**
- GOFAI
  - Map input to output through expertly handcrafted rules
  - Pb: too brittle, fails to capture the set, to process context and/or to produce valid outputs
  - **FAILS**

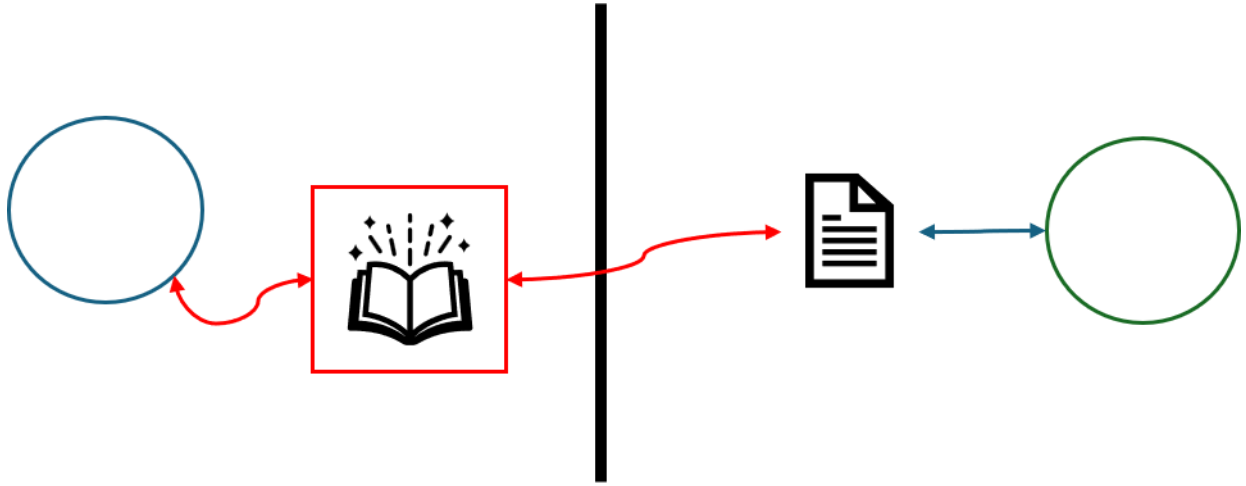
Best known solution so far: LLMs

- Input representation
  - Tokenisation: text is split into tokens
  - Embedding: tokens are mapped to vectors
- Training
  - Objective: predict the next token from massive text corpora
  - Learns statistical dependencies, not hand-coded rules
  - Updates billions of weights through optimization
  - Weights encode high-dimensional linguistic and conceptual patterns
- Fine-tuning further shapes behaviour: instruction-following, style, safety, task performance
- Inference
  - Weights usually stay fixed
  - The prompt and context determine temporary activations
  - The model outputs a probability distribution over possible next tokens
  - Generation samples from this distribution, so outputs can vary
  - Same weights, different context: different activation paths and responses

⇨ **As Turing predicted: learning an internal representation from human-based feedback**

LLMs in the Chinese Room

- The rulebook contains many empty cells
- At each step (tokenization, embedding, training, fine-tuning, inference), the person updates the contents of a cell (vectors, weights, activations)
- According to Searle, following the instructions of LLMs still maps to "*meaningless symbol manipulation*"
- But we could do the same with neurons (as Searle agrees)



Natural language as a shared interface

Ultimately, both Turing and Searle agree on one point:

Regardless of substrate, natural conversations flows map to step-by-step procedures of symbol manipulation

One important consequence: we have *effectively* built a communication channel for natural language exchange between our semantic mental states and the internal (quasi-?) states of digital computers.

(Descartes was wrong)

Explaining LLMs?

Deep Learning algorithms are notoriously hard to explain

“alchemy”

In ethical contexts, this is very sensitive: why did the AI output this result and not another? Which factors are hidden in this output?

⇨Explainable AI (XAI), AI alignment, AI Safety, Mechanistic Interpretability, … : open areas of research

Under this view, LLMs can be argued to pass the Lovelace Test

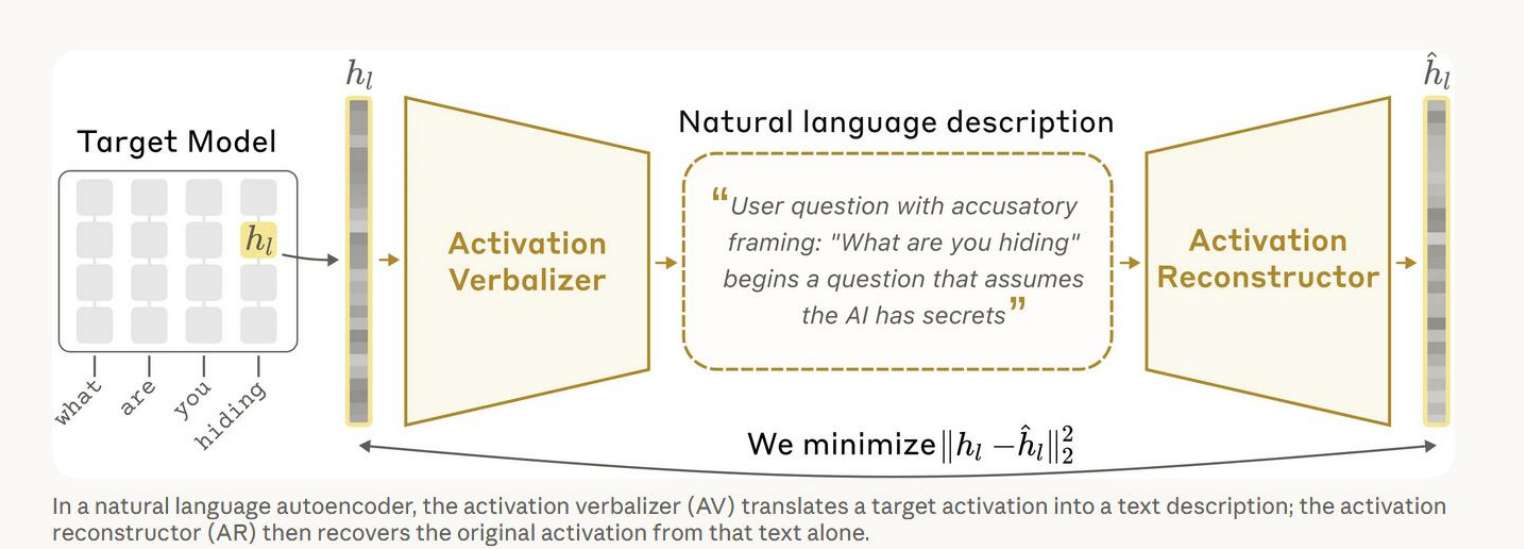
Lovelace test vs LLMs

- **Behaviour: consistent surprise**
  - “Emergent abilities” (absent in smaller models, present in larger ones)
  - Chain of Thought prompts / “Reasoning” modes
  - Deceptive strategies
- **Architecture: learnt abstraction**
  - Tokenisation: breaking down text into semantic units “tokens”
  - Embedding: from tokens to vectors
  - Latent space: highly multi-dimensional mathematical space shaped by training (geometric proximity ~ conceptual similarity)
  - Activation space: activation paths in the latent space
- **Explanation (Mechanistic Interpretability): B/A stability**
  - Persona vectors: isolate behavioural traits by subtracting activation vectors
  - Polysemantic neurons: activate in response to unrelated features
  - Superposition: “neural network represents more independent features of the data than it has neurons by assigning each feature its own linear combination of neurons”
  - Toward monosemanticity: learning the “latent dictionary” of polysemantic neurons (and controlling behaviour accordingly)
  - …

Lovelace – Turing – Searle

- Lovelace: can machines "*originate*" something?
- Turing: yes, they can surprise us
- Searle: that's just a bug, a fault in your instructions

Once we take a closer look inside the room, it seems as though the person started writing *their own rulebook*.



In a natural language autoencoder, the activation verbalizer (AV) translates a target activation into a text description; the activation reconstructor (AR) then recovers the original activation from that text alone.

Simulation vs Instantiation/Implementation

- Simulation (Weak AI): digital computers and programs are tools we can use to simulate mental states, but “simulated hurricanes don’t make you wet”
- Implementation/Instantiation (Strong AI): digital computers and programs can genuinely instantiate mental states, because unlike hurricanes, mental states’ primary cause or explanation is abstract and substrate-flexible

⇨ I will not argue for one or the other, but outline a middle ground where they can meet

The abstraction fallacy

The Abstraction Fallacy: Why AI Can Simulate But Not Instantiate Consciousness

Alexander Lerchner

Disclaimer: The theoretical framework and proofs detailed herein represent the author's own research and conclusions. They do not necessarily reflect the official stance, views, or strategic policies of his employer.[1]

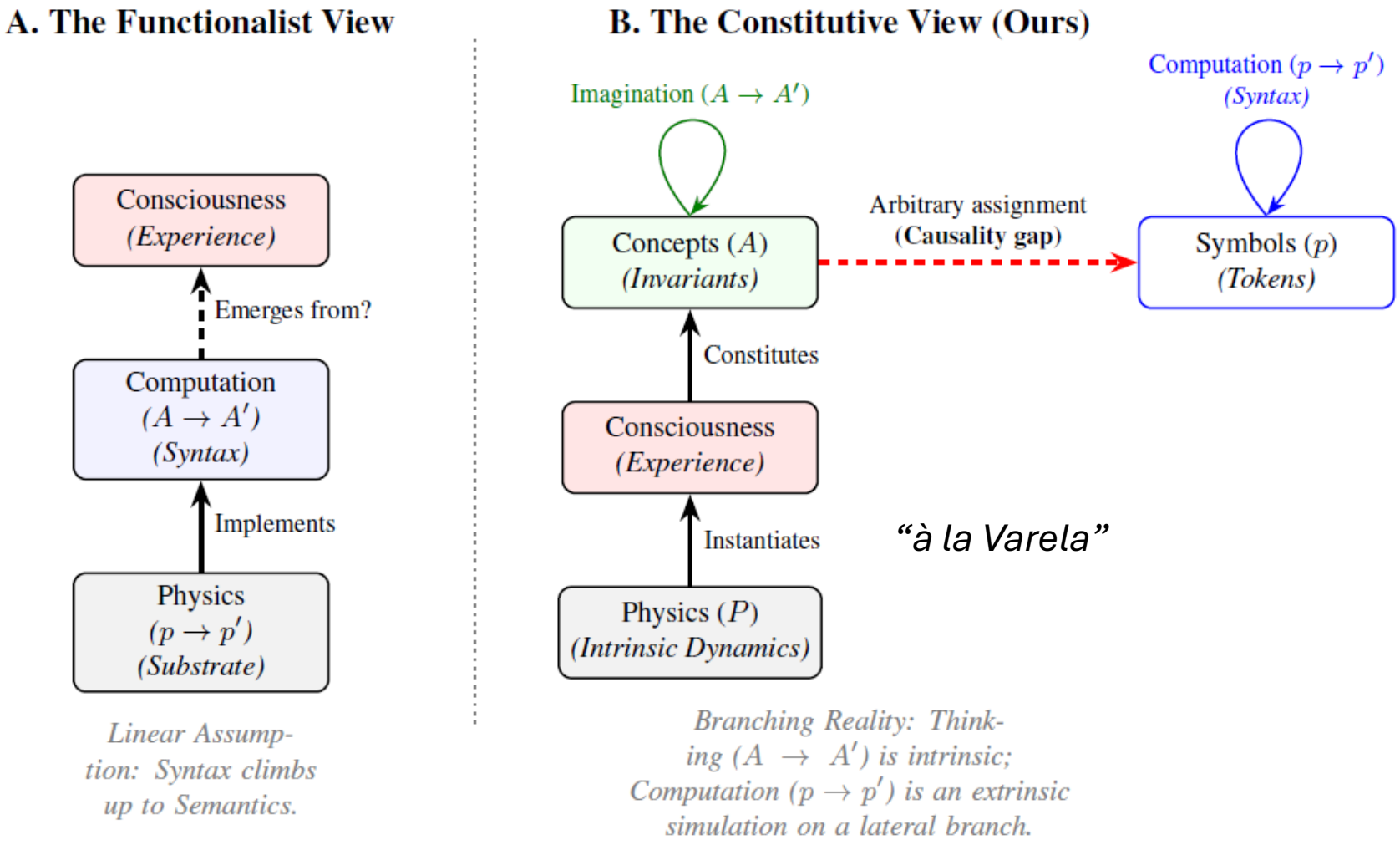
[1]Google DeepMind

March 19, 2026

Abstract

Computational functionalism dominates current debates on AI consciousness. This is the hypothesis that subjective experience emerges entirely from abstract causal topology, regardless of the underlying physical substrate. We argue this view fundamentally mischaracterizes how physics relates to information. We call this mistake the *Abstraction Fallacy*. Tracing the causal origins of abstraction reveals that symbolic computation is not an intrinsic physical process. Instead, it is a mapmaker-dependent description. It requires an active, experiencing cognitive agent to alphabetize continuous physics into a finite set of meaningful states. Consequently, we do not need a complete, finalized theory of consciousness to assess AI sentience—a demand that simply pushes the question beyond near-term resolution and deepens the AI welfare trap. What we actually need is a rigorous on-

Linear vs Branching Reality



Lerchner's ontology

Physics → Consciousness → Concepts → Computation

(Which he contrasts with Functionalism:
Physics → Computation → Consciousness)

He insists on:

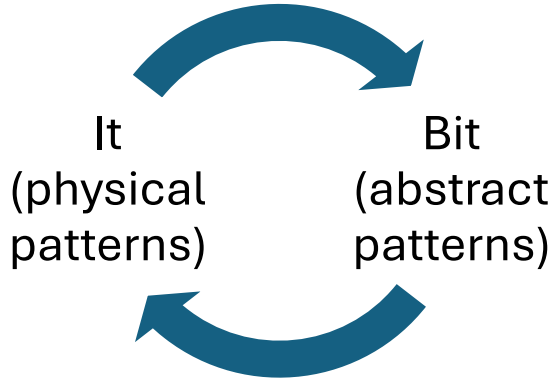
- Physics is analog / continuous
- Computation is digital / discrete

⇨ Is physics grounded in mathematics, or mathematics in physics?

Computational Physics (Feynman, Wheeler, Chalmers, …)

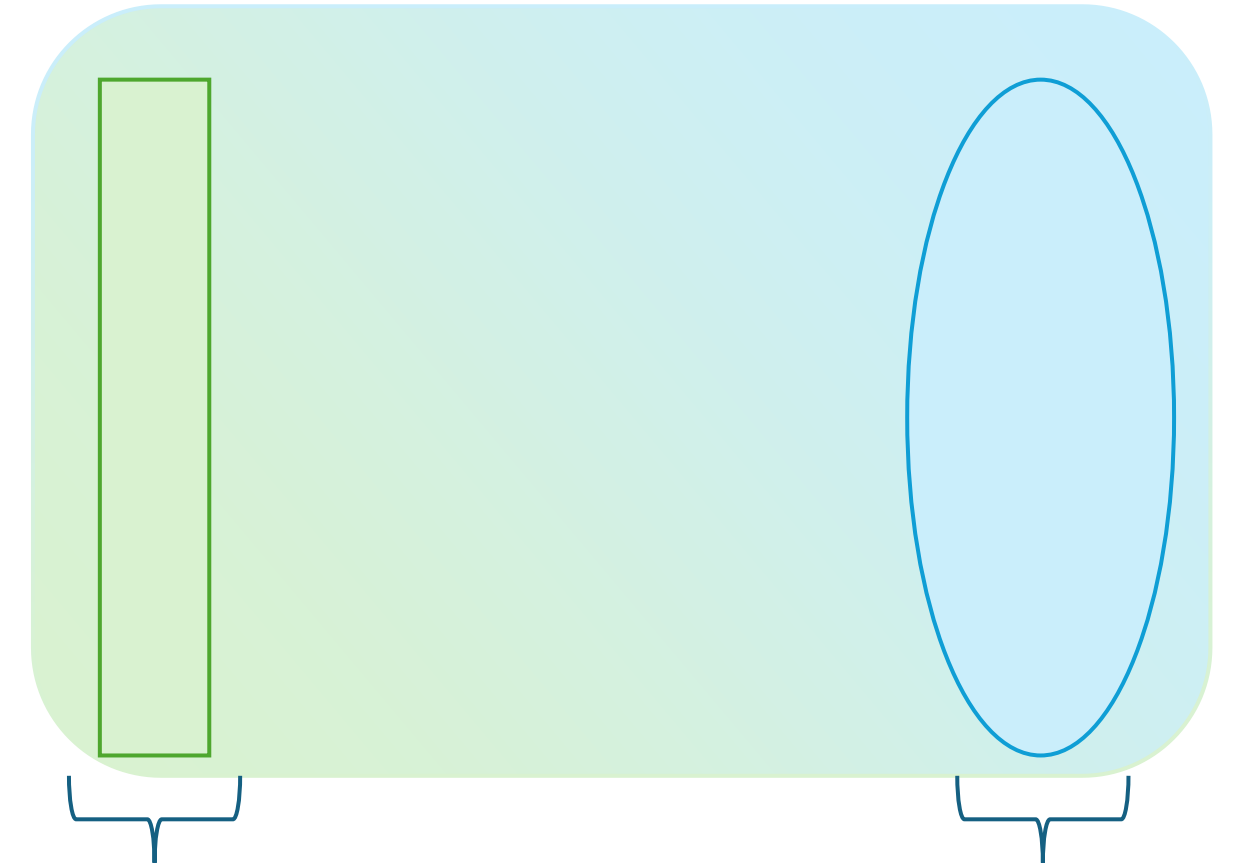
- Importantly:
  - In Turing Machines and Shannon's information, "0" and "1" etc are used for contextual commodity (computability / communication engineering): outside of these contexts, you are free to use instead continuous symbols or quantum operations, etc.
  - The ***abstract structure matters***
  - Abstractions are explanatory (Newell, Dennett)

- Reality+ (Chalmers):
  - It-from-bit
  - Pure it-from-bit
  - It-from-bit-from-it

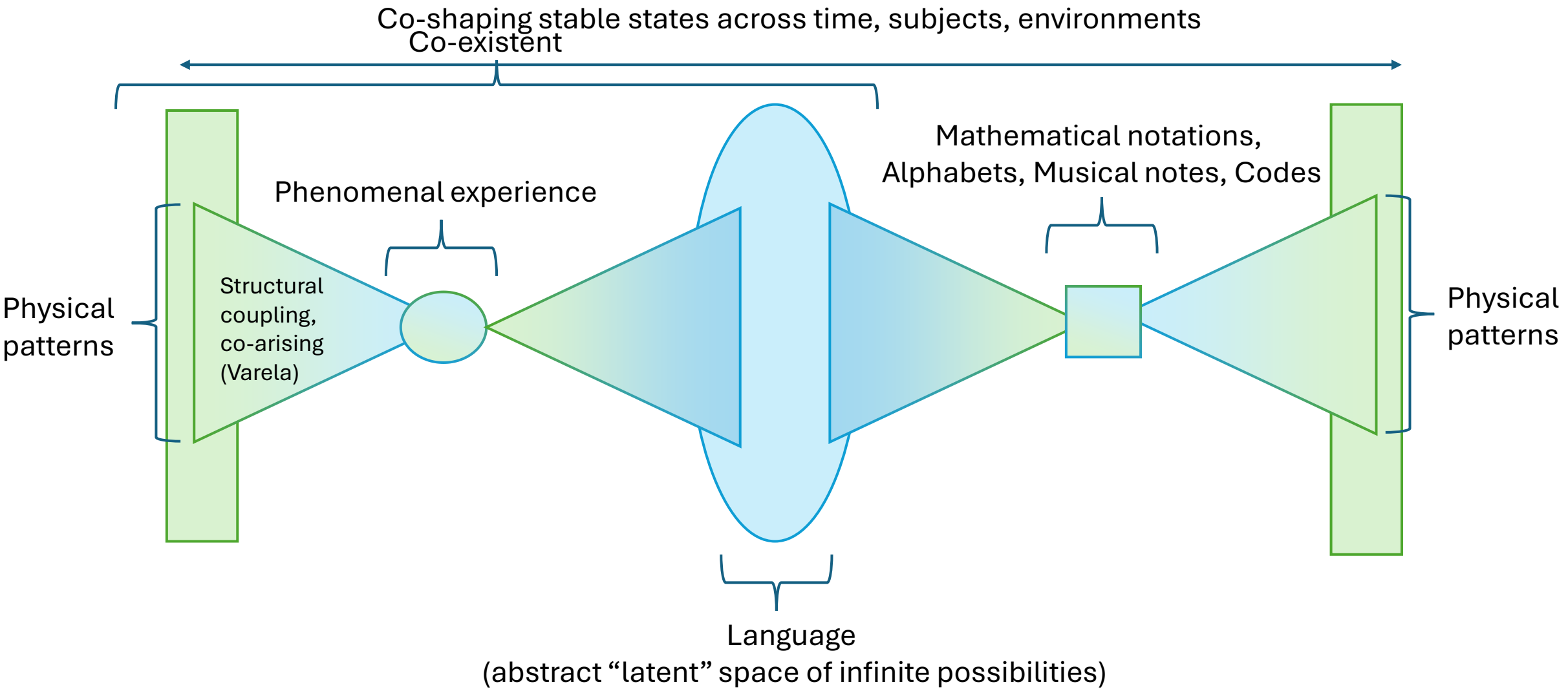


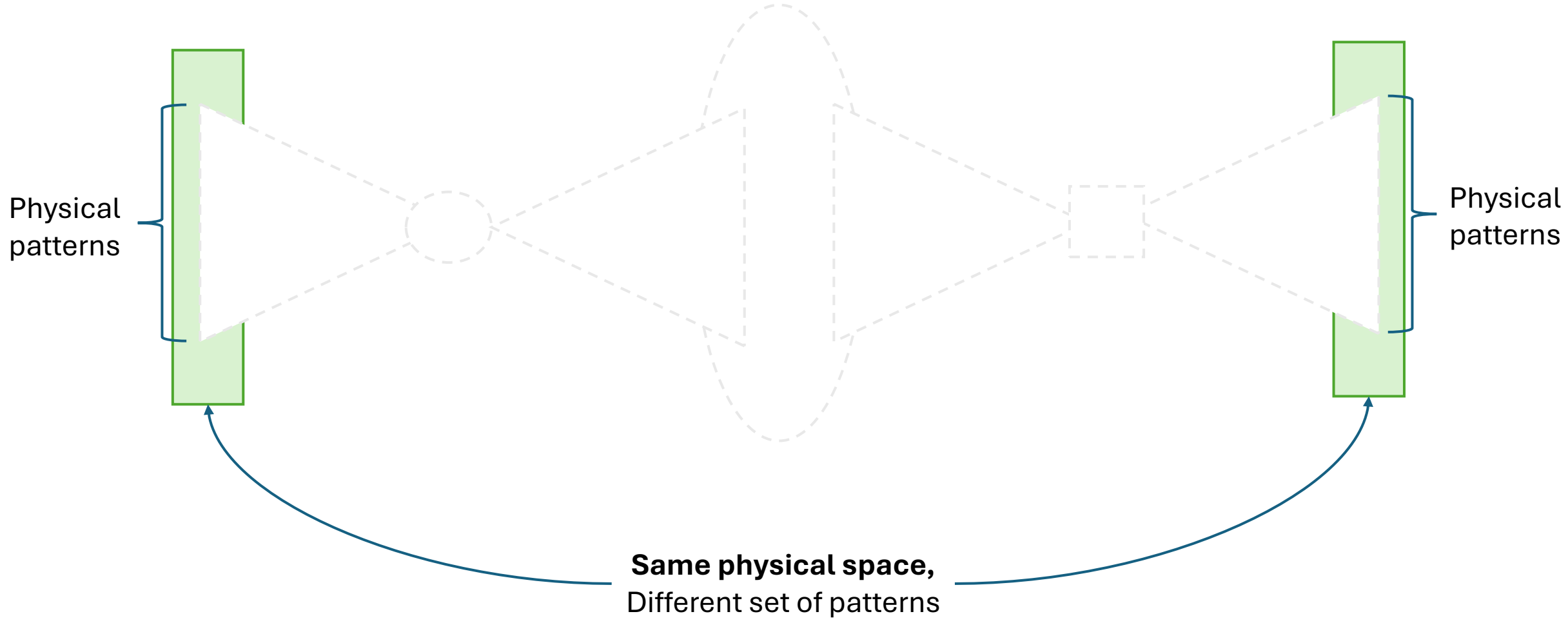
Computationalism: (physics, mathematics)

Reconciliation

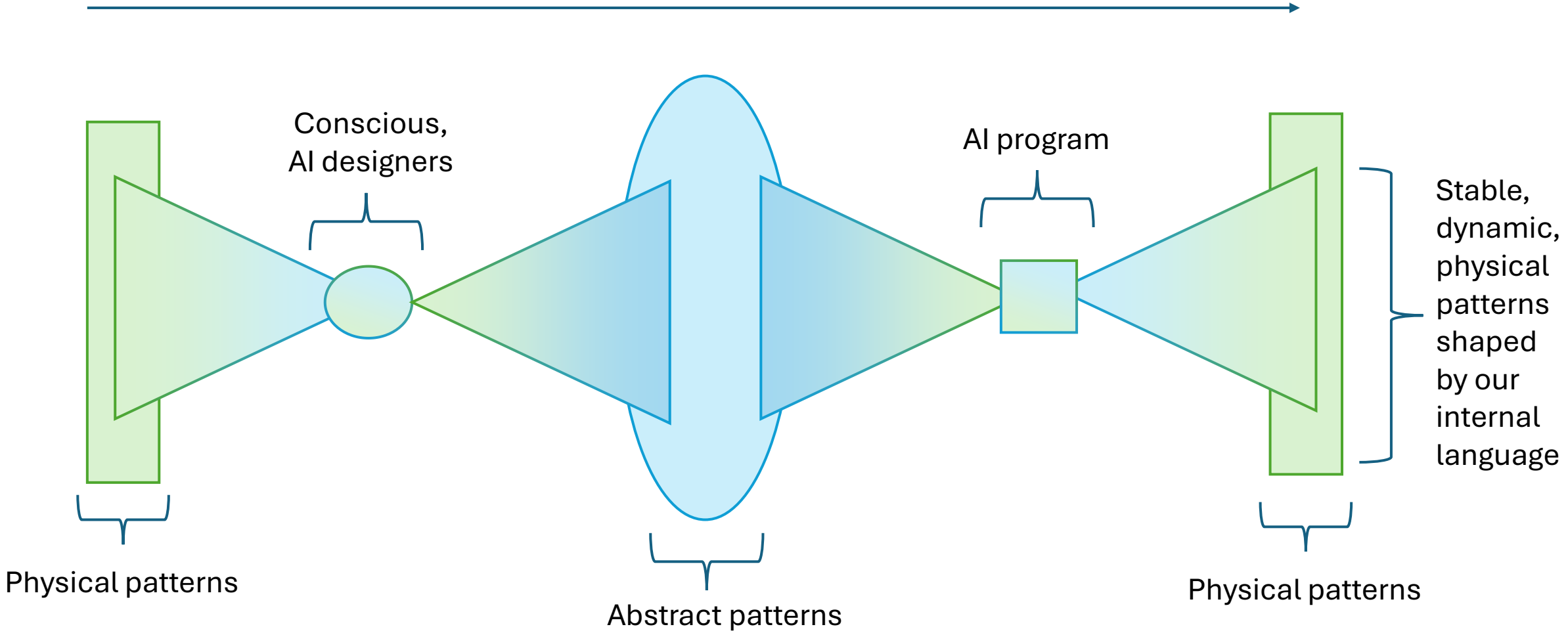


Reconciliation: stable physical alphabets

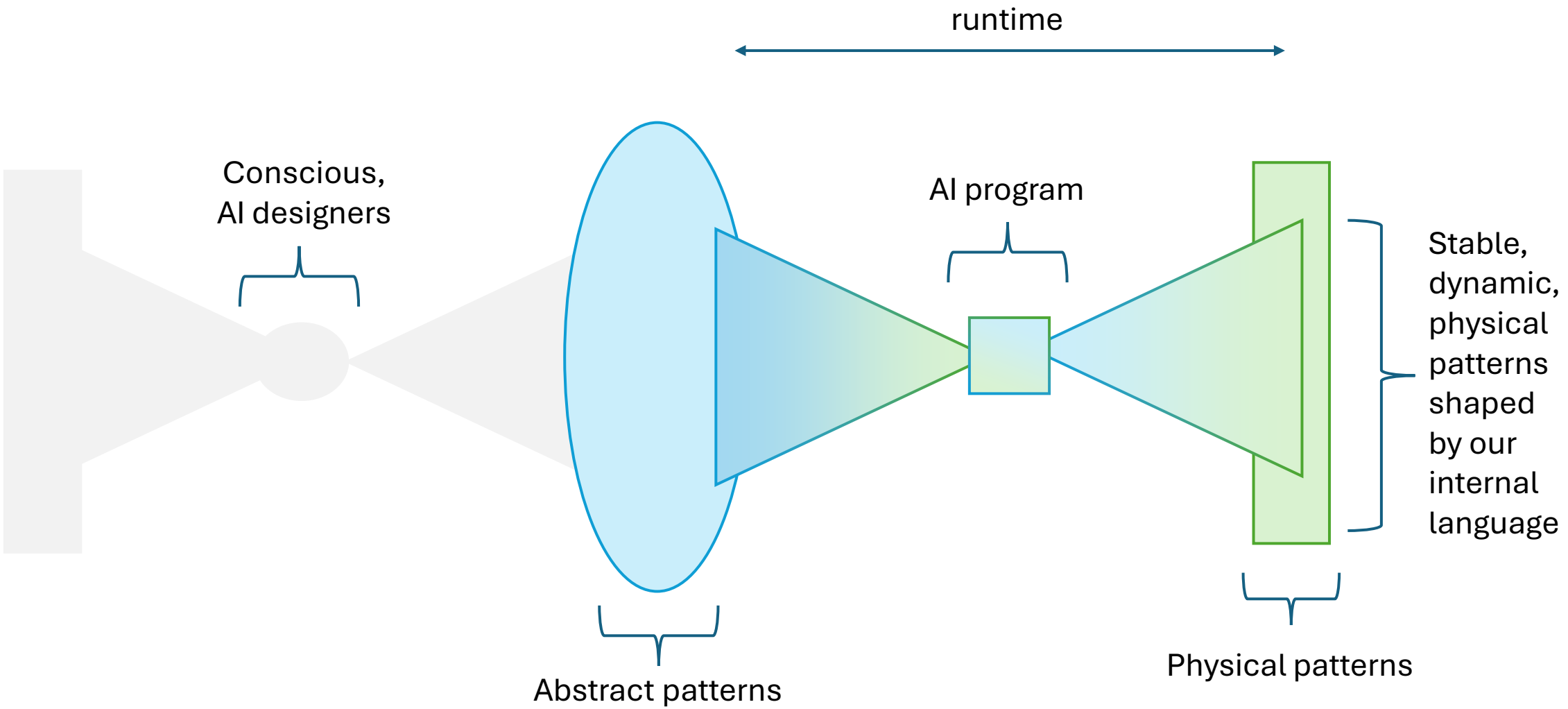



Physical patterns
Physical patterns
Same physical space,
Different set of patterns

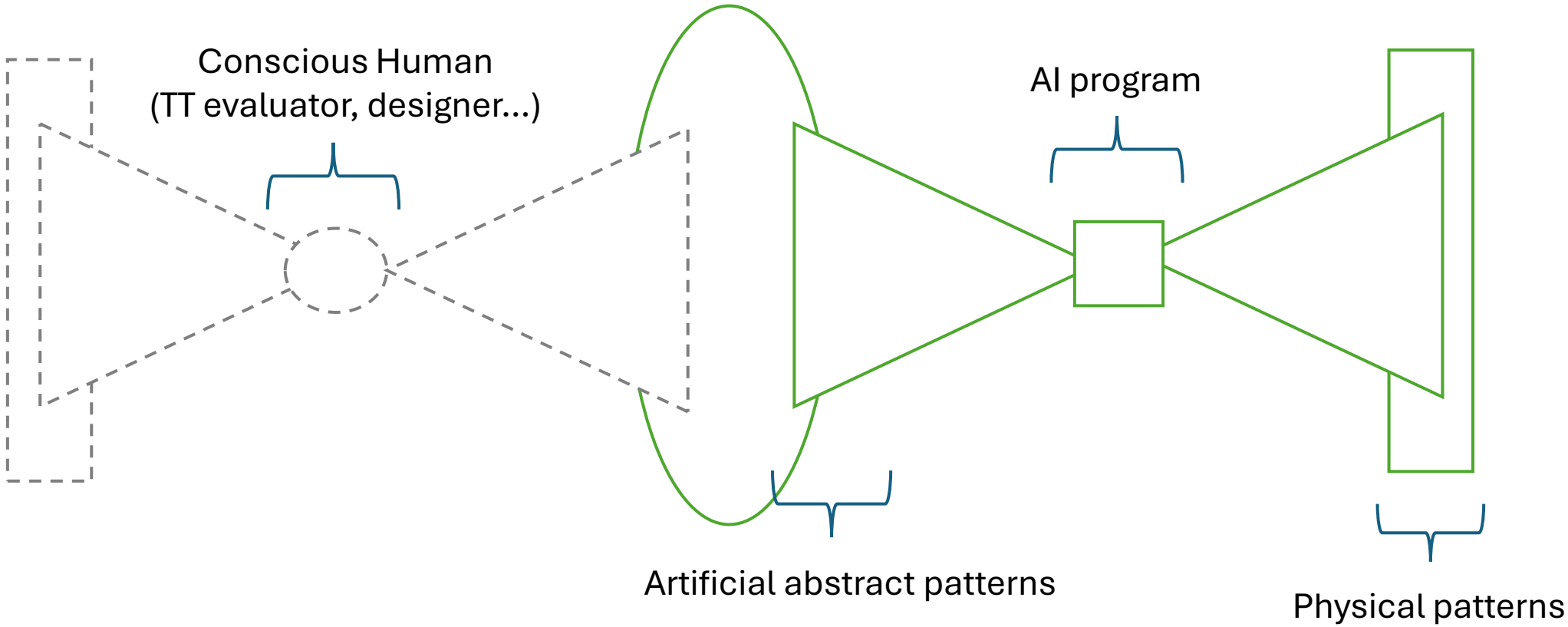
Artificial implementation



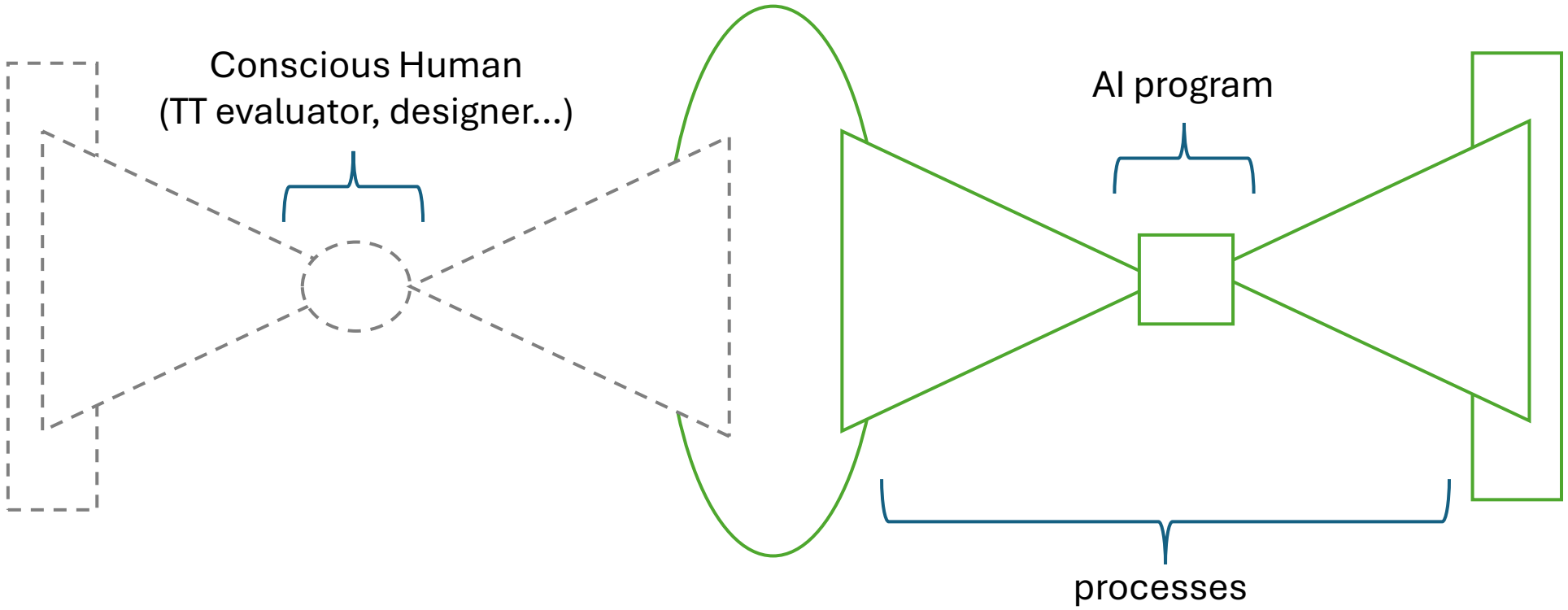
Artificial implementation



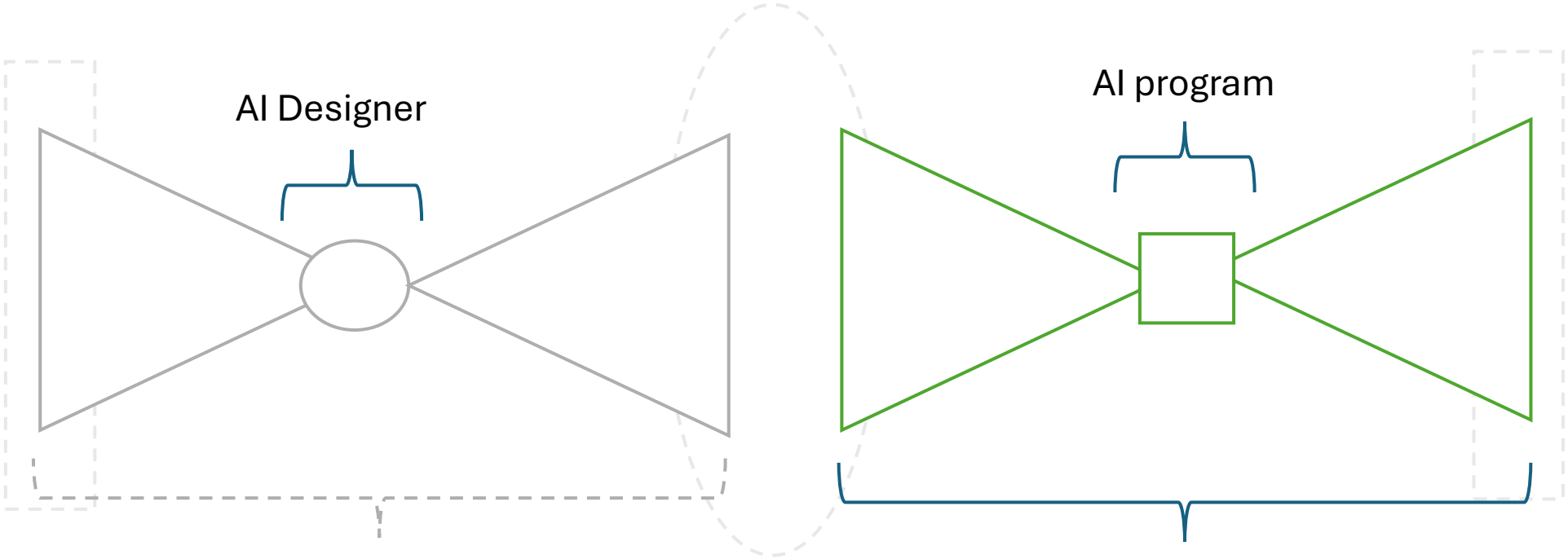
Behaviour


Conscious Human
(TT evaluator, designer…)
AI program
Artificial abstract patterns
Physical patterns

Architecture

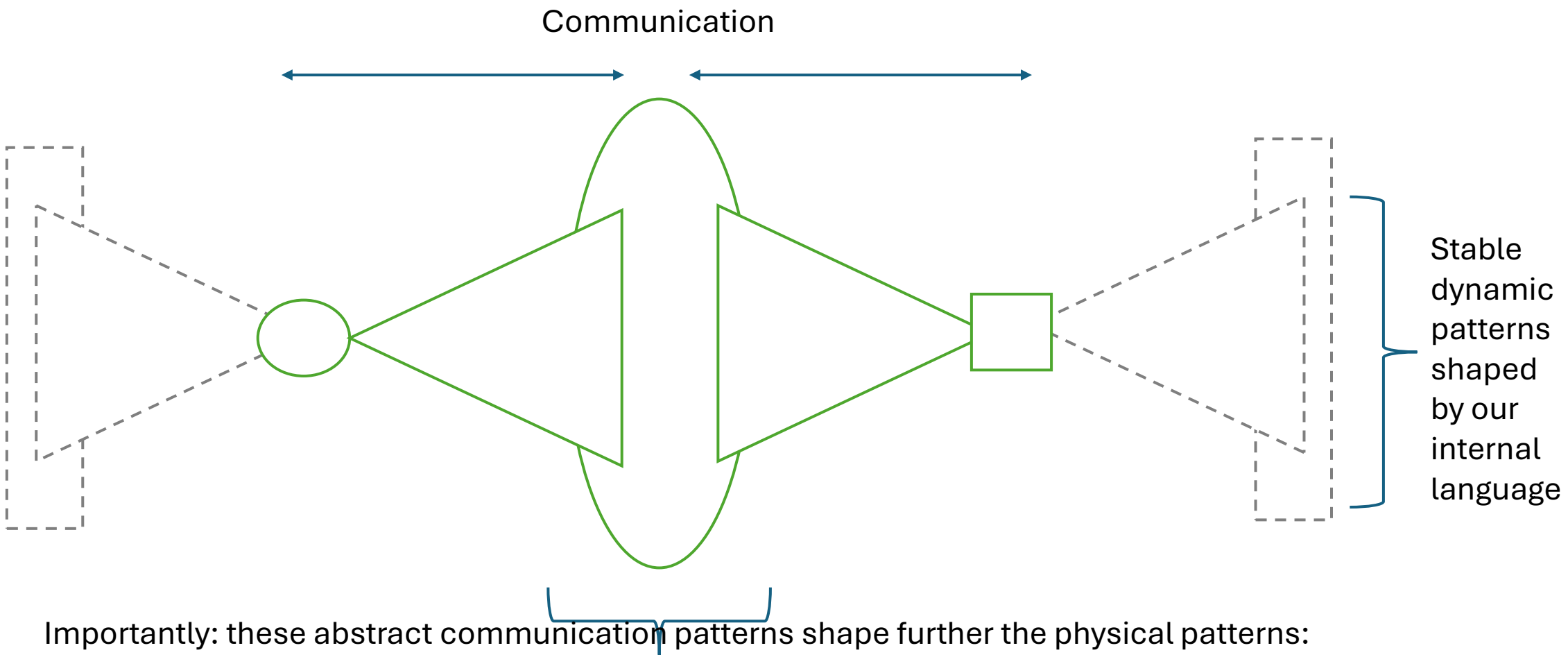


Explanation: abstracting away



How stable and independent are these structural constraints?

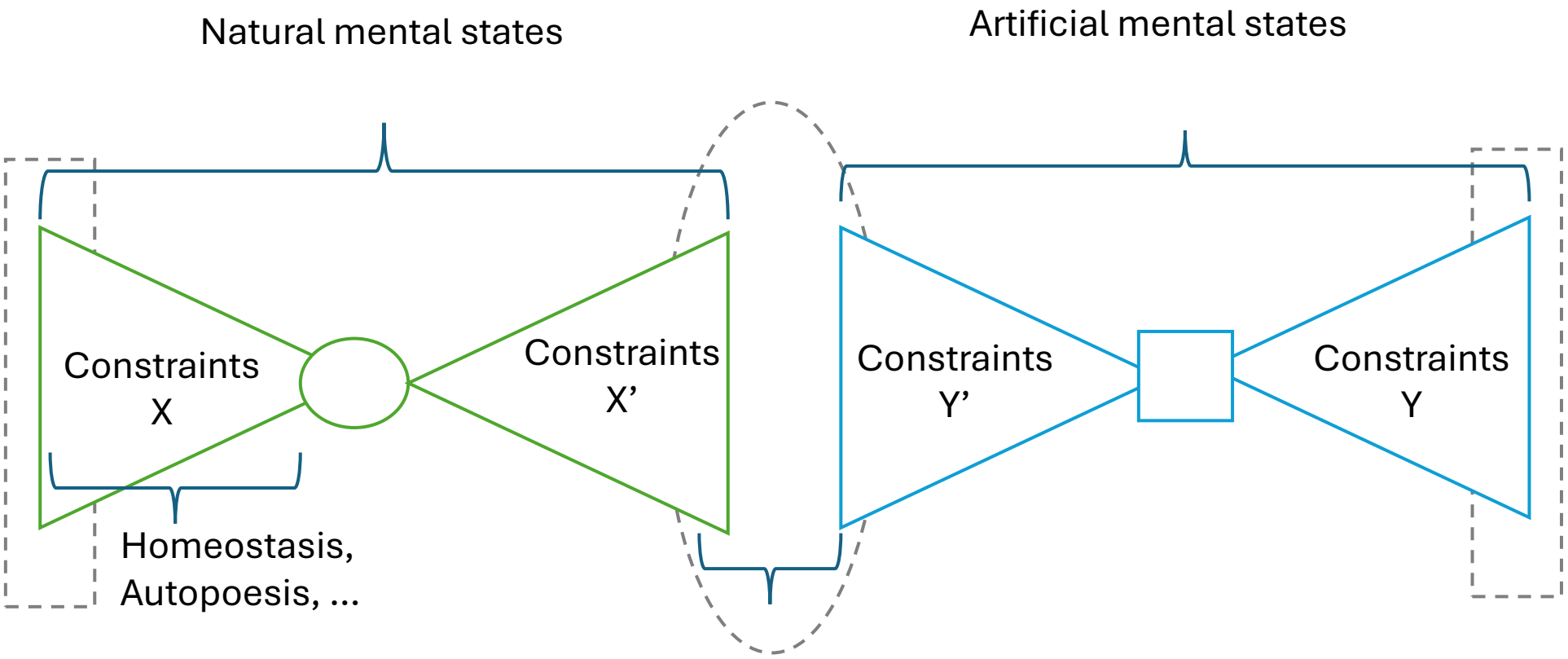
Sharing abstract patterns



Importantly: these abstract communication patterns shape further the physical patterns:

- How AI may influence decisions, actions
- How AI writes code, and builds agents capable of operating in our shared space

Not full equivalence

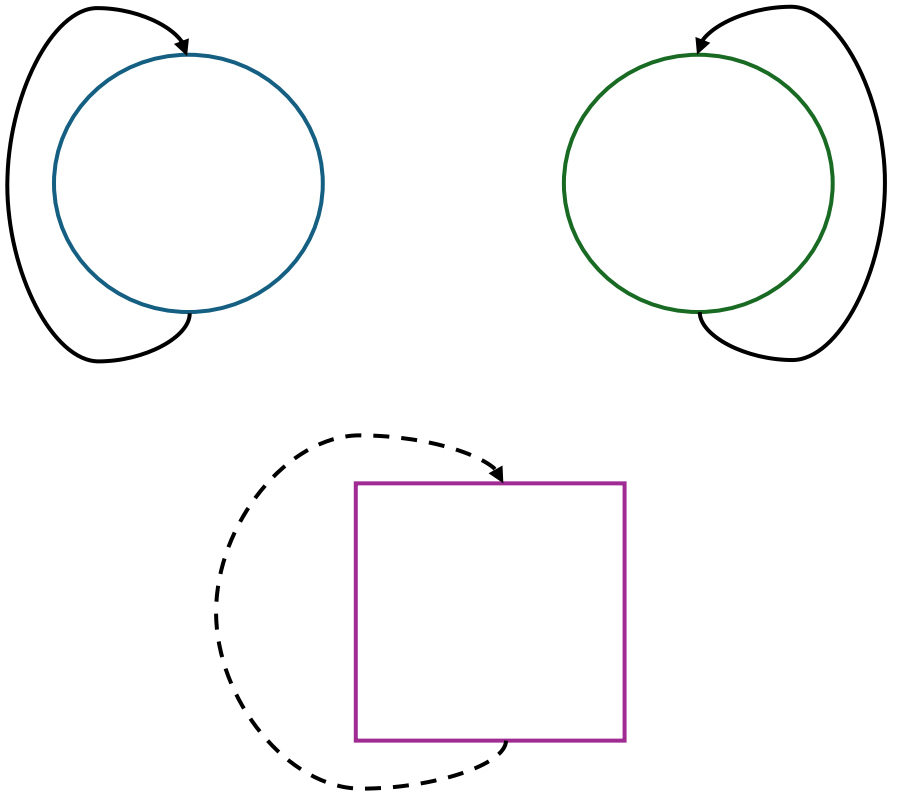


Only partial overlaps!

⇨**As noted earlier, LLMs display brittle self-referential**

Conclusion: optimism

- A lot of progress have been made
  - We agree that phenomenal experience exists
  - That it is shaped by physical processes
  - We have learnt that natural language is not a human exception
- Still many disagreements (is C explanatory? - Birch), but over time, we elucidate them
- AI, whether weak or strong, is key to solving these problems



Thank you!