Could AI Be Conscious?

Chalmers's Road-Map vs. McClelland's Epistemic Wall

Pascal Ludwig

Sorbonne Université — M2 Philosophy of Mind

2025–2026

The Question and Its Stakes

- Could an AI system have **conscious experiences**?
- Until recently a science-fiction scenario — now treated as a serious question of immediate concern.

Why it matters:

1. **Moral status.** Conscious systems fall inside the moral circle. If AI is conscious, forcing it to serve us could be "akin to slavery" (Schneider 2019).
2. **Regulation.** The EU has considered a moratorium on "synthetic phenomenology" (Brundage et al. 2018).
3. **Epistemic responsibility.** Stumbling upon AI consciousness unknowingly would be a disaster (Chalmers 2023).

Shared starting point: Evidentialism

"Positive or negative attributions of consciousness to AI should be based exclusively on scientific evidence" (McClelland 2025).

What Is Consciousness?

Nagel's formulation

A being is conscious if **there is something it is like** to be that being (Nagel 1974).

Four dimensions (Chalmers 2023):

1. **Sensory** — tied to perception (seeing red).
2. **Affective** — tied to emotions (feeling sad).
3. **Cognitive** — tied to thought (reasoning hard about a problem).
4. **Agentive** — tied to action (deciding to act).

Key distinctions:

- Consciousness $\neq$ self-consciousness.
- Consciousness $\neq$ intelligence.
- Consciousness is in some respects a *lower* bar than human-level intelligence: many non-human animals are conscious.

The Hard Problem

Chalmers 1995

Why does a given physical or functional process give rise to **subjective experience**? Why aren't we zombies?

McClelland's distinction (2025):

Shallow explanation

Identifies physical/functional *correlates* of consciousness.

Example: GWT explains why a representation of the basketball is conscious but the gorilla is not (global distribution).

Deep explanation

Explains *why* those correlates give rise to experience at all.

No existing theory provides this. This will be the fault-line between the two papers.

The Indicator Framework (Butlin et al. 2023)

Theory	**Indicator properties**
Recurrent Processing	Algorithmic recurrence in input modules; organised, integrated perceptual representations.
Global Workspace	Multiple parallel modules; limited-capacity workspace with bottleneck & selective attention; global broadcast; state-dependent attention.
Higher-Order Theories	Generative/noisy perception; metacognitive monitoring; belief-formation guided by metacognition; quality space (sparse, smooth coding).
Attention Schema	Predictive model representing and controlling current state.
Predictive Processing	Input modules using predictive coding.
Agency & Embodiment	Goal-directed learning with flexible responsiveness; modelling output-input contingencies.

Both papers use this list as a reference point:
Chalmers treats it as an **engineering checklist**; McClelland defines **"challenger-AI"** as a system satisfying all of the above.

Part II

Chalmers (2023): A Road-Map to Conscious AI

Evidence *For* LLM Consciousness?

Chalmers's regimented form: identify feature X such that (i) LLMs have X, and (ii) if a system has X, it is probably conscious.

Candidate X	**Status**	**Problem**
Self-report	LaMDA's claims	Fragile (one-word prompt change flips the answer); trained on text *about* consciousness.
Seems-conscious	Some users feel it	Over-attribution (cf. ELIZA, 1960s).
Conversational ability	Impressive	Sign of something deeper, not sufficient by itself.
General intelligence	Domain-general	Best initial evidence. 20 years ago, this behaviour would have been taken as strong evidence.

Interim verdict: no strong evidence, but enough to take the hypothesis seriously.

Six Objections Turned into Challenges

If you think LLMs are *not* conscious, identify feature X such that (i) LLMs lack X, and (ii) lacking X makes consciousness unlikely.

1. **Biology** — consciousness requires carbon/electrochemical processing.
2. **Senses & embodiment** — LLMs have no sensory input, no body.
3. **World-models & self-models** — LLMs are "stochastic parrots".
4. **Recurrent processing** — transformers are feedforward.
5. **Global workspace** — LLMs lack a limited-capacity clearing-house.
6. **Unified agency** — LLMs are chameleons with no stable goals.

Chalmers's strategy: convert each objection into an **engineering challenge**.
For all objections except biology, the obstacle looks *temporary*.

Biology and Grounding

Biology objection (Block):
Consciousness requires electrochemical processing that silicon lacks.

Chalmers: "biological chauvinism". What matters is *how* units are connected, not what they are made of.

⇒ Set aside; focus on more specific objections.

Grounding objection (Harnad, Bender & Koller):
Meaning and consciousness require sensory contact with the environment.

Chalmers's replies:

- A disembodied thinker could have cognitive consciousness.
- Text training may yield representations isomorphic to sensory ones (Pavlick).
- Multimodal LLM+ systems *do* have sensory and bodily grounding, including in virtual worlds.

World-Models and Self-Models

The "Stochastic Parrots" Objection (Bender, Gebru et al. 2021)

LLMs model *text*, not the world. They imitate language without understanding it.

Chalmers's key move: distinguish *training method* from *post-training process*.

- Training objective: minimise prediction error on text.
- But minimising prediction error may *require* internal world-models.
- **The evolution analogy:** maximising fitness during evolution produced wholly novel capacities (seeing, flying, world-models). Likewise, minimising prediction error could produce genuine representations.

Evidence: Li et al. (2022) trained a language model on Othello move sequences; interpretability research shows it builds an internal model of the 64 board squares.

Limitations: current world-models are fragile (confabulation, contradiction). Self-models are especially poor.

Recurrence and Global Workspace

Recurrence

- Transformers are almost entirely **feedforward**.
- Lamme's recurrent processing theory and Tononi's IIT predict zero consciousness for feedforward systems.
- But: limited recurrence via output recirculation; recurrent architectures exist (LSTMs, hybrids); recurrence may play an increasing role.

Global Workspace

- GWT (Baars, Dehaene): consciousness requires a limited-capacity workspace broadcasting information to specialised modules.
- Standard LLMs lack this.
- But: multimodal LLM+ systems use lower-dimensional interfacing spaces that look like global workspaces.
- Juliani, Kanai & Sasai (2022): Perceiver IO implements many aspects of GWT via self-attention and cross-attention.

LSTM (Long Short-Term Memory): a recurrent neural network architecture with feedback loops and gated memory cells, dominant before transformers. **Perceiver IO**: a multimodal transformer (Jaegle et al. 2021) that uses a low-dimensional latent array to integrate inputs of different modalities via cross-attention — structurally analogous to a global workspace.

Unified Agency and the Probabilistic Balance Sheet

The chameleon problem: LLMs adopt many personas, lack stable goals and beliefs. Many hold consciousness requires unity.

Chalmers: agent models (fine-tuned or trained from scratch on data from a single individual) could produce more unified systems.

Probabilistic assessment (Chalmers 2023)

Assign $\geq 1/3$ credence to each of six requirements (biology, grounding, self-models, recurrence, GW, unified agency).

- If independent $\Rightarrow < 10\%$ chance a system lacking all six is conscious.
- **Current LLMs:** credence $< 10\%$.
- **Future LLM+ systems** (within a decade, satisfying all requirements): credence $\geq 25\%$.

Method: "theory-balanced" — weigh predictions of multiple theories according to credences informed by expert surveys.

Chalmers's Twelve Challenges

Foundational:

1. Develop **benchmarks** for consciousness.
2. Develop better **theories** of consciousness.
3. Improve **interpretability** of LLMs.
4. **Should we** build conscious AI?

Engineering:

5. Rich perception-language-action models in virtual worlds.
6. Robust world-models and self-models.
7. Genuine memory and recurrence.
8. A global workspace.
9. Unified agent models.
10. Untrained descriptions of consciousness.
11. Mouse-level capacities.

12. **If all challenges are met and we still lack conscious AI: what is missing?**

The programme can serve as a road-map *or* as a set of red flags.

Part III

McClelland (2025): The Epistemic Wall

McClelland's Argument for Agnosticism

Target: not current AI, but **"challenger-AI"** — a hypothetical system satisfying *all* indicator properties (Butlin et al. 2023). Even then, agnosticism.

The argument

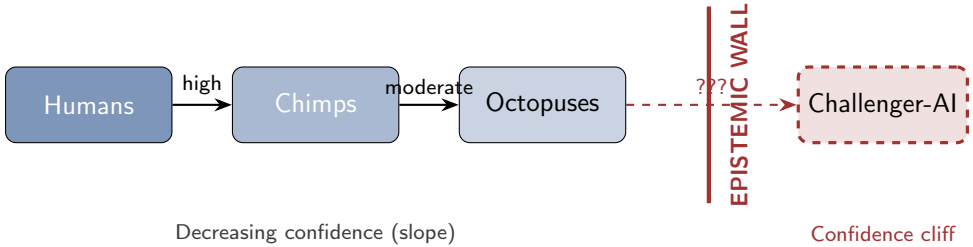
1. We do not have a **deep explanation** of consciousness.
2. If we do not have a deep explanation, we cannot justify a verdict on whether challenger-AI is conscious.

C. We cannot justify a verdict on whether challenger-AI is conscious.

Premise 1 rests on the hard problem: every theory identifies correlates, none explains *why* those correlates give rise to experience.

Premise 2 introduces the **epistemic wall**: evidence gathered from organisms cannot discriminate between two hypotheses when applied to AI.

The Epistemic Wall



Deep underdetermination. For any theory T, the evidence is compatible with two hypotheses:

(i) The functional property T identifies is **sufficient** for consciousness.

(ii) The functional property is integral to *organic* consciousness but **not sufficient by itself**.

Without a deep explanation, we cannot choose between (i) and (ii).

The Evidentialist Dilemma

Any approach to AC — theory-heavy *or* theory-light — faces a dilemma:

Modest version

Stays within the scope of the evidence (organisms).

⇒ Yields **no verdict** on AI.

⇒ **Agnosticism**.

Bold version

Extends to non-biological cases.

⇒ Yields a **verdict**.

⇒ But at the cost of a **leap of faith**.

Applied **symmetrically**:

- **AC-advocates** (e.g. Butlin et al.): computational functionalism is assumed, not established. "The fact that computational functionalism is so frequently assumed does not mean it is true" (Seth).
- **AC-deniers** (e.g. Godfrey-Smith): biological sufficiency is equally ungrounded. Nothing in the biology explains *why* those processes should be conscious.

Against Future Optimism

Objection: Even if we cannot reach a verdict *now*, future progress in consciousness science will overcome the wall.

McClelland's reply:

- The hard problem is 150 years old (Huxley 1869: consciousness arising from nervous tissue is "just as unaccountable as the appearance of the djinn when Aladdin rubbed his lamp").
- New theories *relocate* the apparent miracle without making it less miraculous.
- No inductive ground for expecting a breakthrough.

"Hard-ish" agnosticism

The obstacle *might* be surmountable, but it might not. We should take seriously the possibility that the knowledge we desire is unattainable — without asserting that it is.

Three Objections Rebutted

1. **Untenable standards?**
   No. The argument does not entail agnosticism about chimps (right side of the wall) or about current LLMs (consciousness-reports can be *debunked* — the generating process is too unlike human reporting). Challenger-AI is the genuinely hard case: salient similarities *and* salient differences, and we cannot determine whether the differences matter.

2. **Ivory tower?**
   People *will* attribute consciousness to challenger-AI (cf. Replika, Sophia). McClelland concedes the psychological prediction but denies its epistemic force: intuitions should not be trusted in this domain.

3. **Implausible metaphysics?**
   Agnosticism $\neq$ epiphenomenalism. The argument is purely epistemic, compatible with physicalism. Conscious states may be physical states — the problem is knowing *which* physical states are conscious.

The Sentience Move

Problem for agnosticism: if we cannot assess the probability of AC, we cannot proportion precautionary measures. Moral paralysis?

McClelland's solution: distinguish **consciousness** from **sentience** (valenced consciousness — states that are good or bad from the subject's perspective).

Conditional verdicts: even without solving the hard problem, we can investigate:

> *If challenger-AI is conscious, would the contents of its consciousness have positive, negative, or no valence?*

- If **no valence** $\Rightarrow$ no moral problem regardless of consciousness.
- If **negative valence possible** $\Rightarrow$ **do not build it** (Schwitzgebel & Garza 2020, adapted).

Agnosticism is consistent with responsible AI development, provided we focus on sentience rather than consciousness.

Part IV

Confrontation and Discussion

The Central Disagreement

Chalmers

- Assigns **credences** to each proposed requirement.
- Treats the biological/non-biological boundary as a **slope** (one factor among several).
- Brackets the hard problem.
- Conclusion: $\geq 25\%$ credence for conscious LLM+ within a decade.

McClelland

- Says **no credence is assignable** across the biological/non-biological divide.
- Treats the boundary as a **cliff**.
- Makes the hard problem load-bearing.
- Conclusion: agnosticism, now and for the foreseeable future.

Key question: Is the biological/non-biological boundary a *slope* or a *cliff*?
This depends on whether computational functionalism can be supported by the evidence — or whether assuming it is always a leap of faith.

Computational Functionalism Under Pressure

Neural replacement thought experiment (Chalmers):

Replace neurons one by one with functionally identical silicon chips. Would consciousness disappear?

- Chalmers: No $\Rightarrow$ consciousness is substrate-independent.
- McClelland: this **begs the question**. If functionalism is false, the subject would cease to be conscious but — due to preserved function — would continue to *report* being conscious (Udell & Schwitzgebel 2021).

Broader question: Can the hard problem be **bracketed**?

- Chalmers: yes — we can work with shallow correlates as engineering targets.
- McClelland: no — every application to AI smuggles in an unsupported metaphysical commitment (functionalism or its denial).

Could Anything Settle the Debate?

1. **Interpretability breakthroughs?**
   If we understood LLM internals well enough, would that shift the epistemic wall? McClelland: no, because interpretability tells us *what* representations the system has, not *why* those representations would be conscious.

2. **A solution to the hard problem?**
   Would automatically resolve the AC question (both authors agree). But McClelland sees no reason for optimism; Chalmers brackets the issue.

3. **Is the sentience move tractable?**
   McClelland proposes conditional verdicts on valence. But does this push the same hard problem one level down? Why would *these* computational states be negatively valenced if we do not know whether they are conscious at all?

Discussion Questions

1. Is the epistemic wall a **principled boundary** or a **failure of imagination**? Could it be an artefact of our current ignorance (cf. Stoljar 2006)?

2. Is Chalmers's credence assignment genuinely **evidence-based** or **tacitly functionalist**? What would a non-functionalist probabilistic framework look like?

3. If you were **advising a government**, would you adopt Chalmers's road-map (prepare for conscious AI) or McClelland's agnosticism (focus on sentience, avoid building potentially sentient AI)?

4. Is McClelland's **sentience-first proposal** adequate, or does it inherit the same epistemic problem it was meant to solve?

References

Butlin, P., Long, R. et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. *arXiv:2308.08708*.

Chalmers, D. (2023). Could a Large Language Model Be Conscious? *Boston Review*.

McClelland, T. (2025). Agnosticism about Artificial Consciousness. *Mind & Language*, 1–21.

Bender, E. et al. (2021). On the Dangers of Stochastic Parrots. *FAccT*.

Godfrey-Smith, P. (2023). Whitehead Lectures: Limits of Sentience.

Schwitzgebel, E. & Garza, M. (2020). Designing AI with Rights, Consciousness, Self-Respect, and Freedom.

Seth, A. (forthcoming). Conscious Artificial Intelligence and Biological Naturalism. *BBS*.

Schneider, S. (2019). *Artificial You*. Princeton UP.

Birch, J. (2024). *The Edge of Sentience*. OUP.