AFTER “CONSCIOUSNESS”

What happens if we deliberately set aside the term *“consciousness”* in our thinking about AI and see what grows in the conceptual space it used to occupy? This project treats that as a structured experiment in conceptual engineering: Part I asks whether talk of “consciousness” is distorting philosophy, science, ethics, and public discourse about AI. Part II develops and tests alternative vocabularies drawn from cognitive science, AI practice, and diverse philosophical traditions. Our workshops explore arguments for abandoning “consciousness”, replacement vocabularies, and ask what genuinely new, non-anthropocentric concepts might look like in theory, practice, and governance.

Part I: Why Stop Using “Consciousness”?

You don’t need certainty—just some credence in one or more of these:

1. **“Consciousness” may be a defective concept**
   Failed introductions, endless verbal disputes, culturally parochial.

2. **Illusionism might be right**
   Maybe there is no such property to begin with.

3. **It doesn’t settle what matters**
   The questions that actually matter for AI ethics—Can it deceive? Can it suffer? Should we trust it?—can be investigated without first settling what “consciousness” refers to. We can ask whether an AI system has preferences, goals, or moral status independently of the consciousness question.

4. **Good science has outgrown the label**
   Global workspace, higher-order theories, recurrent processing, predictive processing, IIT, attention schema theory—valuable work regardless of whether it is “really” about consciousness.

→ If any of these are live possibilities, try Part II.

Part II: What Happens If We Just Stop?

The proposal: Remove “consciousness” from the conversation and see what emerges.

1. **Let reductionist programs use their own terms**

    - GWT: broadcast bandwidth, gating mechanisms, downstream integration
    - HOT: meta-representation, tracking accuracy, introspective calibration
    - Recurrent processing: feedback depth, error-correction dynamics
      Same for IIT, Attention Schema Theory, predictive processing, etc.
      Don’t have a competition about who owns the label “consciousness” – that’s a pointless debate. Explore whether these are interesting phenomena, how they relate to each other, and what they explain.
2. **Theorize cognitive and linguistic states directly**
   Use familiar folk-psychological and normative vocabulary:
   speech acts, intentions, representations, agency, welfare-relevant patterns.
   Develop a philosophy of AI mind and language without the “c” detour.

3. **Engineer new AI-specific concepts**
   Rather than forcing AI phenomena into folk-psychological categories (beliefs, desires, experiences) or borrowing human-centric scientific terms, treat this as a genuine conceptual engineering problem: design concepts specifically for AI systems, their architectures, training regimes, and roles in social and political structures.

4. **Explore other traditions**
   Draw on Buddhist, Confucian, Daoist, Vedantic, Indigenous, and other frameworks that foreground different concepts.
   Critical caveat: Some of these might be just as problematic as “consciousness”.

    - Maybe “soul” imports the same mistakes.
    - Maybe some Buddhist concepts are equally defective.
      The point isn’t to endorse these wholesale—it’s to see what happens when we’re not anchored to “c”, and to avoid simply reproducing the same errors in new clothing.

Our workshops will focus on three overarching questions:

1. Are the arguments for abandoning “consciousness” sound?
2. What would replacement vocabularies look like?
3. How do we develop genuinely new concepts without smuggling in human-centric assumptions?

Workshop Framework: Core Questions

1. Methodological Foundations

Should we abandon "consciousness" in AI discourse?

Panel/Talk Topics:

- **Evaluating the Part I arguments**
    - Are these good reasons to stop using "c"?
    - Are there additional reasons not listed?
    - Which is strongest? Which is weakest?
    - Do they compound or conflict?
- **The case for conservatism**
    - What would we lose by abandoning "c"?
    - Are there questions only "c" can ask?
    - Historical parallels: when has concept-abandonment helped or hurt?
- **From Part I to Part II**
    - Do the Part I reasons actually motivate the Part II approach?
    - Or do they motivate something else entirely?
- **Scope questions**
    - Should we abandon "c" just for AI? Or more broadly?
    - What about adjacent terms (sentience, subjectivity, awareness, experience)?

---

2. What Counts as "Replacement"?

Articulating Part II: vocabularies, targets, connections

Panel/Talk Topics:

- **The metaphysics of replacement**
    - What would count as a "replacement vocabulary"?
    - Does it need to target the same phenomena "c" aimed at?
    - Or can it carve up the territory differently?
    - What makes something a replacement vs. a change of subject?
- **Continuity and rupture**
    - What's the relationship between Part II vocabularies and what people were trying to do with "c"?
    - Are we answering the same questions differently?

    - Or asking different questions?
    - How much continuity should we want?
- **Reductionist programs as models**
    - Do GWT, HOT, etc. offer templates for replacement?
    - What can we learn from how they relate (or don't relate) to "c"?
    - When did they succeed or fail by detaching from "c"-talk?
- **Criteria for success**
    - How would we know if Part II is working?
    - What phenomena must a post-"c" framework capture?
    - What problems must it solve?
    - What new tractability should we expect?

---

3. Alien Concepts

What would genuinely new, non-anthropocentric vocabulary look like?

Panel/Talk Topics:

- **AI-specific ontology**
    - What states/properties in large models have no human analog?
    - How do we name phenomena that don't map to folk psychology?
    - Case studies: specific "exotic" AI behaviors.
- **Functionalism without human functions**
    - Can we characterize AI capacities without reference to human capacities?
    - What does agency look like in non-biological, non-evolutionary systems?
    - Goals without drives, preferences without affect—how should we theorize these?
- **Avoiding anthropomorphism**
    - Where does Part II risk smuggling in human-centric assumptions?
    - Which supposedly "neutral" terms carry hidden anthropomorphic commitments?
    - How alien can our concepts get while remaining explanatorily and normatively useful?
- **Measurement and ostension**
    - When we point at AI phenomena, what are we pointing at?
    - How do we stabilize reference to novel properties?
    - What role for operational definitions vs. theoretical identification?

---

4. Non-Western Conceptual Resources

What frameworks foreground different cuts of reality?

Panel/Talk Topics:

- **Mapping the alternatives**
    - Buddhist concepts (dukkha, anattā, skandhas, dependent origination)
    - Confucian frameworks (li, ren, xin, yi)
    - Daoist concepts (wu-wei, ziran, de)
    - Others: Vedantic, Shinto, Indigenous frameworks
    - What do these make salient that “c” obscures?
- **Translation hazards**
    - Risk of importing the same problems as “c”
    - Which non-Western concepts might be equally defective?
    - Which are genuinely orthogonal to Western debates?
    - How to borrow without distorting?
- **Relational vs. intrinsic properties**
    - Traditions that foreground roles, relationships, and processes over intrinsic states
    - How would AI ethics look if we started from relationality?
    - What becomes harder or easier to think?
- **Practical uptake & politics**
    - If we used these frameworks in governance, what changes?
    - Concrete policy implications of conceptual shifts
    - Who benefits, and who is excluded, by different vocabularies?
    - Power and politics of conceptual choice.