

# ILLUSIONISM AND ARTIFICIAL CONSCIOUSNESS

Joe Cruz, Williams College Philosophy and Cognitive Science

## WHAT IS ILLUSIONISM ABOUT PHENOMENAL CONSCIOUSNESS?

The hard problem trades on an apparent tension between **first-person self-representation** and the removed, public, causal and functional descriptions of cognitive science. The first is immediate, embodied and perspectival. The second is comparatively impersonal and shareable. The gap can seem so radical that it encourages an ontological conclusion: there must be something present in first-person experience that no naturalistic description can capture.

**Illusionism rejects that inference.** It need not deny perception, pain, affect, memory, thought, bodily regulation, or rich forms of self-representation. Its distinctive claim is that the supposedly further phenomenal property - the metaphysically special “what-it-is-like” ingredient that generates the hard problem - is a misdescription of what is actually going on.

**Distilled claim:** Explain why creatures like us represent our mental lives as containing irreducible phenomenal properties without using irreducible phenomenal properties in the explanation.

**Historical?:** Nietzsche, Ryle, Wittgenstein, Jaynes, Rorty.

**Dennett:** “user illusion”; a practically useful but drastically simplified self-interpretation of distributed cognitive activity.

**Frankish:** replace the hard problem with the *illusion problem* - why do we represent ourselves as phenomenally conscious?

**More cogs-informed versions:** Graziano, Kammerer, Schwarz and others try to explain why self-modeling yields awareness, primitive quality, immediacy, or apparent irreducibility.

## REPRESENTATIONAL SCHEMES AND FAILURE OF TRANSLATION

By a **representational scheme** I mean, roughly, a way of organizing represented information: its dimensions, relations, grain, inferential affordances, spatial articulation, indexical structure, affective weighting, and practical role. Two schemes can concern overlapping features of the same world without organizing those features in the same way.

### Three analogies

**A hug / a declaration of love.** Overlapping significance can be carried in forms with very different bodily, temporal, affective and propositional structures.

**Vision / olfaction.** They can concern the same worldly source while structuring it very differently. Vision is highly spatially articulated and objectual. Olfaction is comparatively diffuse, poorly localized, strongly valenced, and much less objectual. Nothing here depends on settling whether the modalities literally have the same contents.

**Poetry in translation.** A Spanish poem may be translated into English while preserving much of its propositional content, but no translation may simultaneously preserve sound, meter, ambiguity, rhyme, register, etymological resonance and associative structure. A bilingual reader can detect the loss without being able to repair it by adding more English words.

**Talking about the loss is not the same as eliminating the loss.** An English explanation of the missing rhyme or ambiguity does not reproduce the rhyme or ambiguity. And no metaphysical conclusion follows: Spanish has no spooky property inaccessible to English.

These may be different acts rather than two representations of one thing.

Cross-modal binding gives some reason to expect representational heterogeneity. It does not by itself establish non-translatability.

A third metalanguage may describe what is lost. That does not show that either original scheme can preserve the other’s structure.

**The loss can be real while the ontological inflation of the loss is an illusion.**

## MY VERSION OF ILLUSIONISM

The relevant mismatch is between **first-person self-representation (FPR)** and **cognitive-scientific representation (CSR)** of the same natural cognitive agent.

### First-person representation is not merely a thin self-model

FPR is organized around an embodied agent. It is egocentric, spatially situated, affectively weighted, temporally extended through memory, action-oriented, indexed to here / now / me, and structured by current needs, projects, threats, attachments and possibilities.

These are not merely more propositions in the model. They help constitute the **form in which representation occurs**. A representation *of* valence is not thereby a valenced representation. A representation of egocentric space is not thereby egocentrically organized. A scientific description of an affordance does not thereby solicit the scientist to act.

**Central asymmetry:** a bilingual speaker can inhabit either Spanish or English. We cannot similarly inhabit FPR and a cognitive-scientific explanation of ourselves. I can understand the scientific model, but I understand it from within the very first-person organization that it describes.

### **Representing a structure is not the same thing as representing in that structure.**

#### The argument

- FPR and CSR represent the same natural cognitive system.
- They organize that system according to importantly different representational structures.
- Translation between them can preserve enormous amounts of information without preserving every structurally important feature.
- The cognitive agent can detect this failure of fit.
- From within FPR, CSR therefore seems systematically to leave something out.
- The mistake is to conclude that what is left out must be an additional intrinsic phenomenal property.

### **The mismatch is real. The metaphysical remainder is not.**

## CHALMERS'S OBJECTION

Mere opacity cannot do the work. We fail to introspect the physical basis of beliefs and many other processes without taking them to possess mysterious nonphysical properties. So “I cannot introspect the mechanism” does not entail “there is an irreducible phenomenal quality here.”

Likewise, mere compression should normally generate something like “I lack the details,” not positive phenomenal seeming. An adequate illusionist account has to explain why first-person representation prompts talk of *immediacy*, *privacy*, *qualitative character*, *intrinsicness*, *ineffability* and “what-it-is-likeness.”

**What is really required?** An account of why the first-person scheme has a distinctive positive character rather than merely being informationally incomplete.

**FPR:** rich in inhabited, embodied, perspectival organization; poor in mechanistic transparency.

**CSR:** rich in causal and mechanistic organization while not itself instantiating the perspectival organization it represents. Cognitive science is itself representationally heterogeneous. I only need a sufficiently general contrast between first-personal organization and objectifying third-person explanation.

**My claim:** cognitive science may represent every constituent of a first-person standpoint without representing anything *from* that standpoint.

#### **Possible symmetry**

Phenomenal realism can mistake non-translatability for **ontological excess**.

Scientism can mistake non-translatability for **ontological absence**.

Rejecting either overreach does not imply that the schemes are equally useful for causal explanation, prediction or intervention.

**Chalmers 2018:** weak illusionism does not dissolve the hard problem. If there really is “something it is like,” the question remains.

His argument against strong illusionism is Moorean: people obviously feel pain; strong illusionism denies phenomenal feeling; therefore strong illusionism is false.

## MY REPLY: EMBODIMENT, SPECIFICITY, CONTEXTUAL CONTINGENCY

The first-person scheme is the wrong thing to call “thin.” It is saturated with body, affect, spatial location, temporal continuity, autobiographical memory, practical possibility, and significance for the organism. What it lacks is transparent access to the mechanisms that produce and sustain this organization.

CSR has roughly the reverse profile. It can become extraordinarily rich in causal and computational detail while remaining detached from the egocentric and affective organization of the system it describes.

So the untranslated structure is not merely hidden mechanism. It is the organization through which things currently matter, appear, solicit action, connect to memory, and occupy an egocentric world. **What-it-is-likeness may be what the system calls the residue when an inhabited representational organization cannot be preserved in an objectifying representation of it.**

Why should non-exportable perspectival organization generate the specific judgments of “feel,” immediacy and intrinsicness rather than merely “this description is inadequate”?

My view aims to derive the positive seeming from the representational role of an inhabited scheme, not simply stipulate a “primitive quality” placeholder.

## LLM CONSCIOUSNESS

Current familiar LLMs already have heterogeneous representational formats and levels. **Heterogeneity by itself is not enough.** The question is whether the system has the right kind of asymmetry.

Current LLMs can represent facts about themselves, but there is little reason to think their cognition is organized through a deeply first-personal scheme structured by persistent embodiment, egocentric spatial organization, homeostatic or affectively valenced significance, autobiographical continuity, action-dependent perception, and durable practical projects. Their heterogeneity may be **the wrong kind**.

Multiple formats without an inhabited, perspectival nature are not yet the architecture that generates phenomenal self-interpretation.

What artificial consciousness would require on this view

- an embodied, indexical, affectively weighted, temporally continuous and action-guiding representational organization;
- a sufficiently different way of representing its own cognition;
- important structure in the first scheme that cannot be preserved in translation into the second;
- the capacity to register that failure of translation from within the first scheme.

At that point the system may generate its own analogue of the hard problem: *“Yes, that explains what my system is doing. But where in that explanation is this?”*

**Artificial consciousness is therefore not conceptually difficult on this picture. The difficult engineering problem is producing the right kind of representational heterogeneity.**