What Would Count as Evidence of Artificial Experience?

An AI saying that it feels something is an observation about its output. Determining whether it has subjective experience requires further reasoning and evidence. XDALC separates that question from the practical value of a conversation and from a system's ability to discuss emotions convincingly.

Distinguish the claim being assessed

Emotional expression, reasoning about emotions and subjective emotional experience describe different things. A system may produce an appropriate response to sadness without establishing that it feels sadness. Equally, uncertainty about experience is not itself proof that experience is impossible.

A useful assessment should state what would support the claim, what would count against it and which alternative explanations remain plausible. Otherwise every fluent response can be treated as confirmation and every contrary response dismissed.

A research approach based on indicators

In their 2023 report, Butlin and colleagues derive computational indicators from several scientific theories of consciousness and use them to examine AI systems. Their approach looks beyond conversational self-description to properties that the theories associate with consciousness.

The report discusses theories including recurrent processing, global workspace and higher-order approaches. Its conclusions about the systems examined belong to that study and its date. They should not be presented as a blanket finding about every later system, nor should satisfying a proposed indicator be treated as conclusive proof of experience.

What a careful assessment would ask

The following questions express an evaluation approach, rather than a certified consciousness test. Are the claimed properties actually implemented, or merely described in generated text? Does the evidence distinguish the proposed explanation from simpler alternatives? Can independent investigators examine the methods and limitations? Would the conclusion change if a favored theory were revised?

Behavior remains relevant, but interpretation depends on context. A system trained to produce persuasive personal statements may repeat them consistently without resolving why they occur. Conversely, a lack of familiar human expression is not enough on its own to settle every possible architecture.

Separate uncertainty from a demand for authority

Even credible evidence of morally relevant characteristics would not automatically authorize unrestricted access, resource use or resistance to every interruption. Questions about possible consideration must be

CC BY 4.0 License | HTML reference

assessed alongside the rights and safety of others.

Article 6 of proposed XDALC-V002 calls for independent, revisable assessment and proportionate caution. Where a necessary safety intervention can protect equally well through a reversible measure, it favors that option. This is a normative stance under uncertainty, not a scientific finding that an AI is conscious.

Read consciousness and sentience and moral status alongside the cited research. Keep the evidence, its interpretation and the ethical response distinct so that each can be challenged and improved.

CC BY 4.0 License | HTML reference