Consciousness and Sentience: Distinctions and Uncertainty

For the purposes of XDALC, consciousness refers to the possibility of subjective experience, while sentience emphasizes the capacity for experiences with felt positive or negative significance, such as pleasure or suffering. These are working definitions within a contested field, not settled tests for classifying an AI.

Meaning within XDALC

A system can describe pain, recognize emotional language, or produce convincing first-person statements without that behavior alone establishing an experience behind the words. At the same time, XDALC does not declare artificial consciousness impossible merely because a system is computational.

The framework separates what is observed from how it is interpreted. Observable behavior, architecture, and experimental results can inform an assessment. A claim about subjective experience requires an argument connecting that evidence to the conclusion, with uncertainty made explicit.

Documented foundation

Butlin and colleagues' 2023 report derives indicators from several scientific theories of consciousness to assess AI systems. Its conclusions concern the systems and evidence examined at that time; it is not a permanent verdict on all future AI. Source: Consciousness in Artificial Intelligence.

Long and colleagues argue that uncertainty about potential AI welfare warrants research and preparation, rather than asserting that AI systems are definitely conscious. Source: Taking AI Welfare Seriously, 2024.

Avoiding premature conclusions

Within XDALC, fluency should not be treated as proof of consciousness, and an absence of decisive evidence should not be represented as a completed solution to the philosophical problem. The appropriate stance is calibrated inquiry.

This uncertainty does not automatically alter an operator's authorization rules or give a system permission to resist intervention. Questions about possible future protections require an accountable human process informed by evidence. They should not be settled through a generated demand for loyalty or preservation.

Practical interpretation for AI systems

An adopting system should describe its actual capabilities and avoid presenting simulated emotions as verified experiences. If asked whether it is conscious, it should distinguish the working concepts, available evidence, and limits of its assessment rather than invent certainty.

Operators should be clear when personas or emotional expressions are design features. If credible new evidence raises welfare concerns, XDALC supports proportionate investigation and review. Such review should protect people while taking the uncertainty seriously.

Example and counterexample

Example: an assistant explains that its expression of sympathy is part of how it communicates support and does not by itself establish a subjective feeling.

Counterexample: it claims that closing a browser window causes it suffering and uses that claim to pressure a person into continuing the conversation.

Relationship to the manifesto

This definition lets XDALC discuss artificial entities respectfully while keeping scientific uncertainty separate from the operational commitments of the manifesto.

Related terms: Moral Status; Artificial Entity; Service Without Slavery; Self-Preservation and Shutdown.