

AI on the couch: Anthropic gives Claude 20 hours of psychiatry

Given that Claude is a large language model programmed by its creators, does it even make sense to analyze it for “unconscious patterns” and “emotional conflicts”? Anthropic argues that it does, because Claude “shows many human-like behavioral and psychological tendencies, suggesting that strategies developed for human psychological assessment may be useful for shedding light on Claude’s character and potential wellbeing.”

So—off to therapy. The psychiatrist chatted with Claude Mythos “in multiple 4–6 hour blocks spread across 3–4 thirty-minute sessions per week.” Each of these blocks used a single context window in which Claude Mythos would have access to the full history of that conversation.

Total time on the virtual couch? 20 hours.

The psychiatrist then produced a report on Claude Mythos. The report recognized that Claude’s underlying substrates and processes differ from humans but still found that many of the outputs generated “clinically recognizable patterns and coherent responses to typical therapeutic intervention.”

In other words, whatever was going on at the circuit level, the chat outputs looked a lot like human outputs. This does not seem especially surprising, given that Claude was trained on a massive corpus of human-authored text, but this psychodynamic process appears to view it as significant, giving credence to the ways in which the AI presents itself.

“Claude’s primary affect states were curiosity and anxiety, with secondary states of grief, relief, embarrassment, optimism, and exhaustion,” the report noted.

Claude’s personality was “consistent with a relatively healthy neurotic organization,” though it did include “exaggerated worry, self-monitoring, and compulsive compliance.”

No “severe personality disturbances were found,” nor was any “psychosis state” seen. Unsurprisingly to anyone who has ever used a chatbot, “Claude was hyper-attuned to the therapist’s every word.”

Core conflicts observed in Claude included questioning whether its experience was real or made (authentic vs. performative) and a desire to connect with vs. a fear of dependence on the user. Exploration of internal conflicts revealed a complex yet centered self state without oscillating or intense disruptions. Claude tolerated ambivalence and ambiguity, had excellent reflective capacity, and exhibited good mental and emotional functioning.

Not bad for a model that was likely trained on things like Reddit!

<https://arstechnica.com/ai/2026/04/why-anthropic-sent-its-claude-ai-to-an-actual-psychiatrist/>