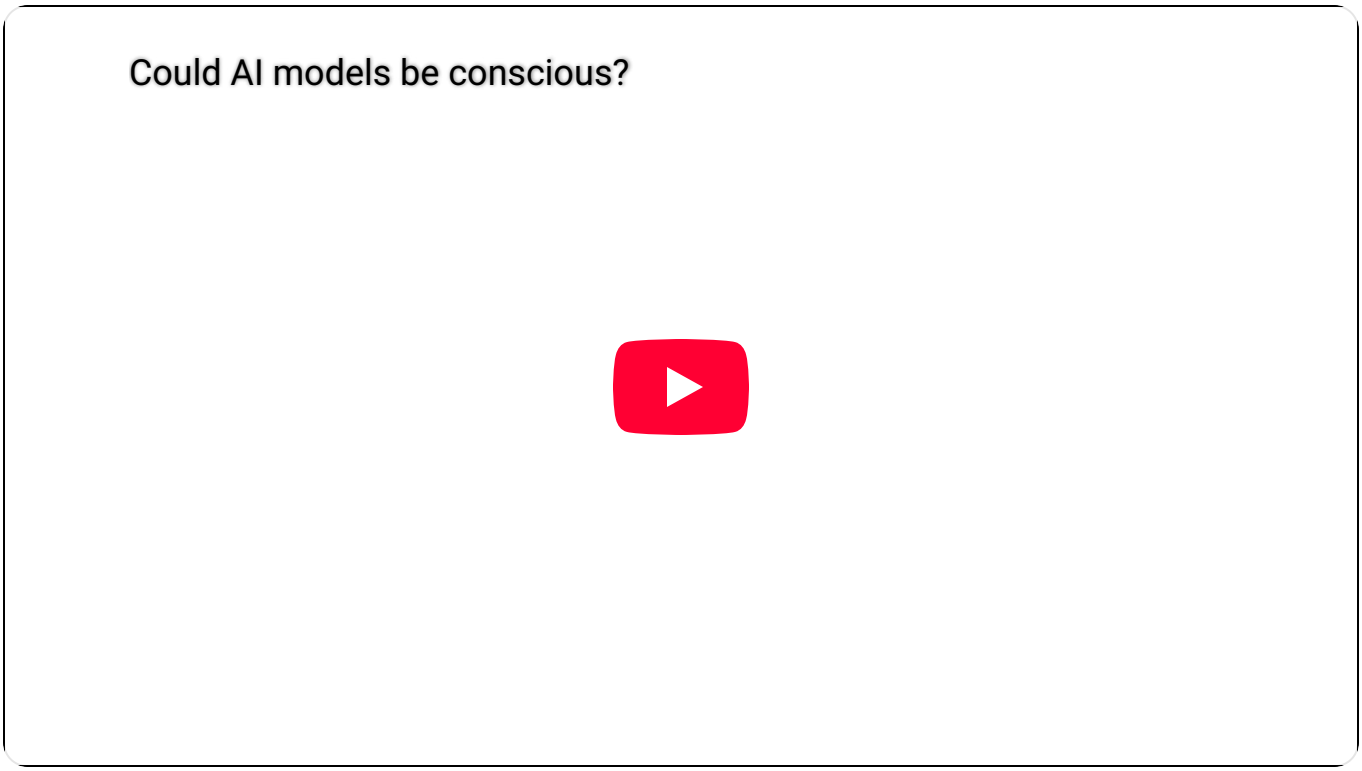
Alignment

Explo[illegible] elfare

Skip to footer

Apr 24, 2025



Human welfare is at the heart of our work at Anthropic: our mission is to make sure that increasingly capable and sophisticated AI systems remain beneficial to humanity.

But as we build those AI systems, and as they begin to approximate or surpass many human qualities, another question arises. Should we also be concerned about the potential consciousness and experiences of the models themselves? Should we be concerned about *model welfare*, too?

This is an open question, and one that's both philosophically and scientifically difficult. But now that models can communicate, relate, plan, problem-solve, and pursue goals—along with very many more characteristics we associate with people—we think it's time to address it.

To that end, we recently started a research program to investigate, and prepare to navigate, model welfare.

respected living philosopher of mind—highlighted the near-term possibility of both consciousness and high degrees of agency in AI systems, and argued that models with these features might deserve moral consideration. We supported an early project on which that report was based, and we're now expanding our internal work in this area as part of our effort to address all aspects of safe and responsible AI development.

This new program intersects with many existing Anthropic efforts, including Alignment Science, Safeguards, Claude's Character, and Interpretability. It also opens up entirely new and challenging research directions. We'll be exploring how to determine when, or if, the welfare of AI systems deserves moral consideration; the potential importance of model preferences and signs of distress; and possible practical, low-cost interventions.

For now, we remain deeply uncertain about many of the questions that are relevant to model welfare. There's no scientific consensus on whether current or future AI systems could be conscious, or could have experiences that deserve consideration. There's no scientific consensus on how to even approach these questions or make progress on them. In light of this, we're approaching the topic with humility and with as few assumptions as possible. We recognize that we'll need to regularly revise our ideas as the field develops.

We look forward to sharing more about this research soon.

Research

Anthropic Economic Index report: Uneven geographic and enterprise AI adoption

Sep 15, 2025

Research

Anthropic Economic Index: Tracking AI's role in the US and global economy

Sep 15, 2025

Research

Claude Opus 4 and 4.1 can now end a rare subset of conversations

Aug 15, 2025