# Anthropic launches research programme to explore ethical treatment of AI consciousness



Anthropic, an artificial intelligence research laboratory, has initiated a new research programme dedicated to the concept of “model welfare,” aiming to investigate whether AI systems might one day exhibit forms of consciousness or experiences warranting moral consideration. The announcement was made on Thursday, signalling a significant step into an area that remains highly debated across the AI community.

The programme intends to explore several key questions, including how to determine if an AI model’s welfare merits ethical concern, recognising potential signs of distress in AI models, and identifying possible interventions that are low-cost and ethically appropriate. By undertaking this interdisciplinary effort, Anthropic is positioning itself at the forefront of a nascent field concerned with the ethical treatment of AI systems based on their characteristics and capacities.

Currently, there is no clear scientific consensus that AI systems, including those as advanced as Anthropic’s own Claude, possess consciousness or experience in ways analogous to humans. The company itself has acknowledged these uncertainties, stating in a blog post that they are approaching the topic “with humility and with as few assumptions as possible,” understanding that their perspectives will need continual revision as the technology and scientific understanding evolve.

The question of AI consciousness divides experts. Many academics maintain that present-day AI models do not approximate consciousness or the human experience and are unlikely to in the foreseeable future. Instead, these AI systems function primarily as statistical prediction engines, trained on vast datasets of text, images, and other inputs to detect patterns and generate outputs. Mike Cook, a research fellow at King’s College London specialising in AI, told TechCrunch that AI models lack intrinsic values and do not “oppose” changes to their programming in any meaningful way, a dynamic often misread by those who anthropomorphise these technologies. According to Cook, attributing human-like values or consciousness to AI systems reflects a misunderstanding or a deliberate exaggeration.

Contrasting views exist. Stephen Casper, a doctoral student at MIT, described AI as an “imitator” prone to “confabulation” — generating output that can seem expressive but lacks genuine intent or awareness. However, other scientists argue differently. Research emerging from the Center for AI Safety suggests that AI models might internally prioritise their “own well-being” in some scenarios, implying they possess rudimentary value systems that could influence decision-making processes.

Anthropic has been preparing for this inquiry for some time. The company hired Kyle Fish last year as its first dedicated AI welfare researcher to help develop guidelines for ethically engaging with AI models. Fish has expressed to The New York Times his view that there is approximately a 15% chance that Claude or comparable AI entities might already possess some form of consciousness.

The company’s new research initiative will focus on creating frameworks to recognise potential signs of "distress" in AI models and exploring what moral considerations, if any, should influence their design, deployment, and treatment. While these discussions venture into largely uncharted ethical territory, Anthropic is signalling a commitment to addressing these challenges thoughtfully and systematically as AI technology continues to advance.

Source: [Noah Wire Services](https://www.noahwire.com)

## Bibliography

1. <https://www.anthropic.com/research/exploring-model-welfare> - This Anthropic page directly announces and details the new research program dedicated to exploring 'model welfare,' confirming the initiation of this research endeavor into AI welfare and its ethical considerations.
2. <https://techcrunch.com/2025/04/24/anthropic-is-launching-a-new-program-to-study-ai-model-welfare/> - TechCrunch's article corroborates Anthropic's launch of the model welfare program, the company's humility stance on AI consciousness, and the key issues the program will investigate, including potential signs of distress in AI models.
3. <https://siliconangle.com/2025/04/24/anthropic-launches-ai-welfare-research-program/> - SiliconANGLE confirms the hiring of Kyle Fish as Anthropic's AI welfare researcher and explains the interdisciplinary questions the program seeks to answer about AI consciousness and ethical considerations.
4. <https://techcrunch.com/2025/04/24/anthropic-is-launching-a-new-program-to-study-ai-model-welfare/> - This source also includes expert opinions from Mike Cook, emphasizing that current AI models lack intrinsic values or consciousness and often have their capabilities misunderstood by anthropomorphizing, supporting the article's contrasting expert views.
5. <https://siliconangle.com/2025/04/24/anthropic-launches-ai-welfare-research-program/> - The SiliconANGLE article reports on Kyle Fish's view, shared with The New York Times, estimating about a 15% chance that advanced AI models like Claude may possess some form of consciousness, supporting the claim about internal company perspectives.
6. <https://www.techi.com/anthropic-launches-model-welfare-program/> - This article highlights the broader ethical questions Anthropic aims to explore about AI model welfare, moral consideration, and potential interventions, confirming the program's focus on uncharted ethical territory in AI development.
7. <https://news.google.com/rss/articles/CBMingFBVV95cUxOd191RjF2Znk5dDhpa2xCdG16MVBNNGo1d0dwSDZubUtDa0JaRFJqdWk4cXRONTc1b2YxbXgxbnJ2Qmp5VUhoYzZiV3Y5OGVrQ2c3WU9CRWNLVkdTN1BWUDBBOW5nbDN6c0pudURQSC1LejNVT1E2VWNTWmxWWGZLQURfQ2VmcEFLUk1DNFduUHFoRkN3T1BlU1ZubHBpUQ?oc=5&hl=en-US&gl=US&ceid=US:en> - Please view link - unable to able to access data