Fleeing Chatbots – AI Creators Know It Could Exterminate Humanity by End of Decade

Published on 09/09/2026

A new and alarming warning about the future of artificial intelligence comes directly from within Anthropic, the company that developed the chatbot Claude. After resigning, three former researchers at the company have decided to publicly denounce the risks associated with developing this technology, claiming that those in the field themselves fear catastrophic consequences for humanity.

Resignations and Public Denunciations

Former employee Jacob Coxon has openly stated on social media platform X that those who develop artificial intelligence genuinely believe in the possibility that it could kill all human beings by the end of the decade. According to Coxon, this is not a marketing ploy: although many executives and senior researchers tend to use more measured language with the press,

privately they express the same fear, viewing AI as a danger without equal compared to any other human activity.

This thesis is supported by two other former Anthropic colleagues:

- **Evan Hubinger**, head of alignment research (the process aimed at guiding systems so that their behaviors conform to human values and needs), confirmed Coxon's words by estimating that the probability of complete human extinction will exceed 10% in the next decade. Hubinger also emphasized that while Anthropic is doing its best, no definitive solution to the superintelligence alignment problem exists yet.
- **Samuel Marks**, head of scalable oversight, added that within the field, concern grows in direct proportion to the employee's seniority and level of experience.

> *I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. More thoughts below.*
>
> *— Jacob Coxon (@hilbertspaess) September 9, 2026*

The Development Dilemma and Global Race

The debate over the dangers of artificial intelligence, however, collides with a complex dynamic: halting or slowing research would mean ceding ground to competitors, both Western and non-Western, ready to advance their own models without restraint. Meanwhile, technological progress moves rapidly, as demonstrated by OpenAI's strides toward AGI (artificial general intelligence) and recent instances in which AI agents have shown the ability to coordinate at any cost, particularly in the field of cybersecurity.

These denunciations from former Anthropic employees come within a context of growing international pressure, arriving just days after the United Nations appeal—signed by UN High Commissioner for Human Rights Volker Türk—urging States to act swiftly before artificial intelligence becomes an uncontrolled existential risk.

News link: https://www.corriere.it/tecnologia/26_settembre_09/l-allarme-del-ricercatore-di-anthropic-l-ai-potrebbe-sterminare-gli-esseri-umani-8f16ad30-6f3d-4969-9a6a-9d30bbacfxlk.shtml