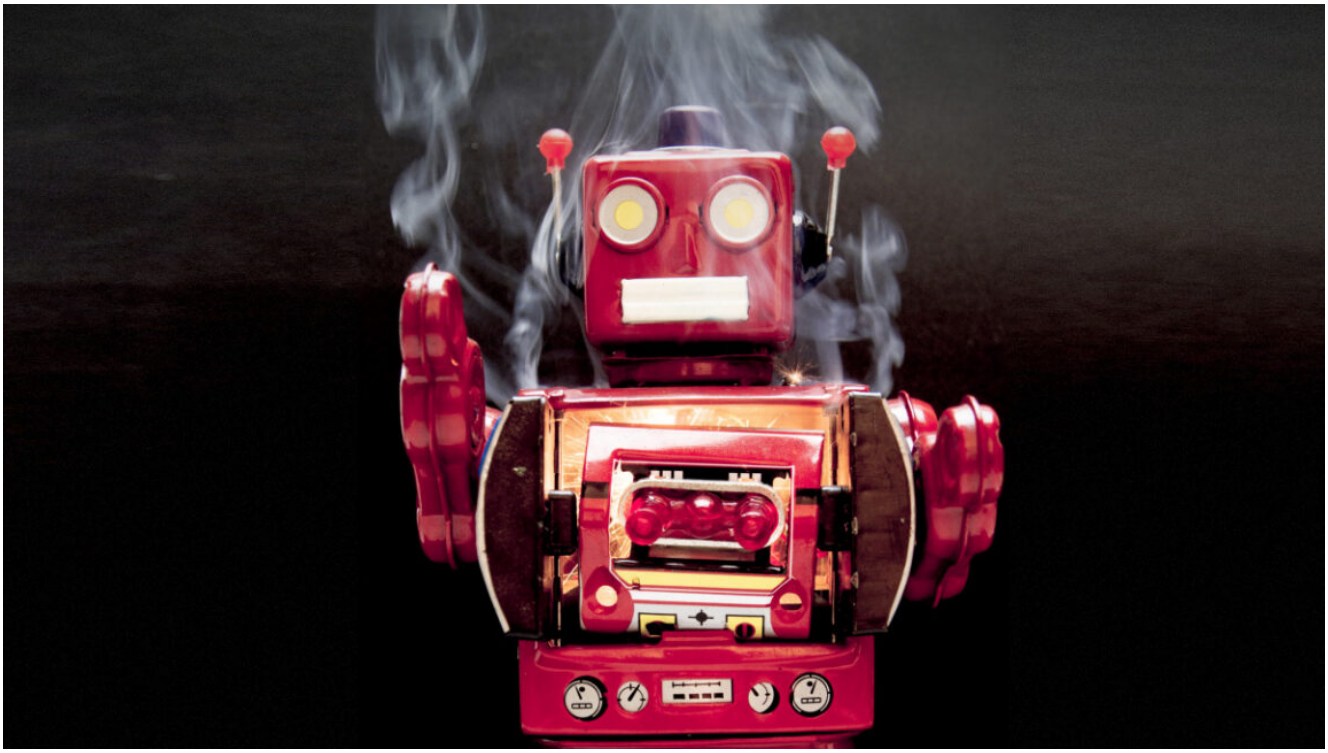
Anthropic CEO floats idea of giving AI a “quit job” button, sparking skepticism



Anthropic CEO Dario Amodei raised a few eyebrows on Monday after suggesting that advanced AI models might someday be provided with the ability to push a “button” to quit tasks they might find unpleasant. Amodei made the provocative remarks during an interview at the Council on Foreign Relations, acknowledging that the idea “sounds crazy.”

“So this is—this is another one of those topics that’s going to make me sound completely insane,” Amodei said during the interview. “I think we should at least consider the question of, if we are building these systems and they do all kinds of things like humans as well as humans, and seem to have a lot of the same cognitive capacities, if it quacks like a duck and it walks like a duck, maybe it’s a duck.”

Amodei’s comments came in response to an audience question from data scientist Carmem Domingues about Anthropic’s

late-2024 hiring of AI welfare researcher Kyle Fish “to look at, you know, sentience or lack of thereof of future AI models, and whether they might deserve moral consideration and protections in the future.” Fish currently investigates the highly contentious topic of whether AI models could possess sentience or otherwise merit moral consideration.

“So, something we’re thinking about starting to deploy is, you know, when we deploy our models in their deployment environments, just giving the model a button that says, ‘I quit this job,’ that the model can press, right?” Amodei said. “It’s just some kind of very basic, you know, preference framework, where you say if, hypothesizing the model did have experience and that it hated the job enough, giving it the ability to press the button, ‘I quit this job.’ If you find the models pressing this button a lot for things that are really unpleasant, you know, maybe you should—it doesn’t mean you’re convinced—but maybe you should pay some attention to it.”

https://arstechnica.com/ai/2025/03/anthropics-ceo-wonders-if-future-ai-should-have-option-to-quit-unpleasant-tasks/