Stanford eCorner

Balancing AI Priorities

23-02-2024

URL: https://stvp.stanford.edu/clips/balancing-ai-priorities

Daniela Amodei, president and co-founder of Anthropic, tells a story from Anthropic's early days to illustrate how companies building AI tools have to balance helpfulness, honesty, and harmlessness. The company is always trying to improve all three, she says, but different use cases may require different balances.

Transcript

- We wanna be responsible with large language models, 00:00:05,700 how do we approach that in the most honest and transparent way? - So I'll start with kind of a funny story here 00:00:12,360 to sort of illustrate what you're talking about, which is in sort of early days of training Claude, we were really experimenting with sort of trading off some of these like H's, right? This like helpful, honest, harmless.. You can have a perfectly harmless model if you wanted.. It would just not be very helpful, right? You know, we would sort of ask Claude like, who was the first president of the United States? And it would be like, "I cannot answer that question and I'm also very concerned about your wellbeing.. And like, here is a link to like, you know, harm prevention website," and you're just like, "Claude, I'm fine.. I promise I'm fine." So there is sort of this, there is sort of this like chart of like sort of intersection, right? Where you can say, okay, like do we want Claude to like risk like a little bit more helpfulness for like a little bit more honesty, right? Or a little more harmfulness or however you might describe it.. And of course what we're always trying to do is sort of raise the watermark on all three.. But at the end of the day, depending on the application, you might also want to sort of fine tune the model or train a different model for certain use cases, right? You can imagine that for an educational tool for like, you know, five to eight-year-olds, you might want a very, very harmless version of Claude, even if it's like a little bit less helpful.. Whereas if you're using Claude to do, for example, you know, trust and safety detection work, which is an application that many of our customers use Claude for so that humans don't have to look through harmful content, you actually want Claude to be able to read and identify and understand very harmful or upsetting things and sort of filter them out.. - That's right...