Stanford eCorner

Constitutional AI

Daniela Amodei, *Anthropic*

23-02-2024

URL: https://ecorner.stanford.edu/clips/constitutional-ai/

Daniela Amodei, president and co-founder of Anthropic, explains her company's approach to building trustworthy, reliable, and safe generative AI systems. Anthropic's AI tool, Claude, is governed by a constitution composed of several documents, including the UN's Universal Declaration of Human Rights.

Transcript

- Tell us more about Anthropic.. 00:00:06,960 Tell us about what the foundational, philosophical ethic is of the company.. You've innovated some really interesting things on how to make a responsive, explainable, steerable AI.. Maybe we'll just start with an open book.. Like, share with the audience what Anthropic is all about.. - Sure.. 00:00:24,120 So really, Anthropic is aiming to build transformative general AI systems that are safe, reliable, steerable, ethical.. And that's kind of, like, it's a lofty goal, right? So much of, I think, why we were founded was this desire to build generative AI tools and systems and products that people could use while feeling really, really confident that what we were putting on the market was trustworthy, reliable, and safe.. And that's taken a number of different forms, you know, over the past few years.. But I think specifically, we have this sort of belief that incorporating these technical safety streams into training of the model, really, from day one is the best way to ensure that when they actually are in a product form and getting into the hands of customers, that they're gonna be safe..

- I love how your team has thought about using 00:01:20,120 the UN Human Rights Declaration as a starting point for how to really moderate content and moderate the algorithm's development.. Can you speak to that a little bit? - Definitely.. 00:01:33,930 So this kind of idea that we came up with, our brilliant research team came up with is, this idea of giving Claude, which is our, you know, generative AI tool, something we call a constitution.. It's called Constitutional AI.. And to just sort of go back a little bit to kind of how we got there, the way that you sort of used to, and by used to, I mean like three years ago or a year and a half ago.. - Or like, three days ago.. 00:01:56,460 - You used to train, you know, these models 00:01:59,310 to make sure, hey, they're not sort of saying nasty things, or they're not replicating bad inherent biases, was you would do this technique called reinforcement, learning from human feedback, which we also kind of co-worked on when we were at OpenAI.. And reinforcement learning from human feedback is just essentially a way of, like, giving the model, like, grades, (laughing) right? It's like, A plus, you said something nice.. Like, D minus, you said something mean.. And that was reasonably effective in changing some behavior of the models..

But what we found was that a lot of sort of subtler things about how models respond, react, things that they're sort of deeply believing are much harder to train out in those sort of individual cases.. And so we had this idea of giving the model kind of a broader constitution, in the way that you would in a society, to basically say: What are ways of behaving and

engaging that are good for, you know, humans, right, for people and that don't perpetuate some of the problematic things that might be in training data? And so so much of, you know, what we saw with Constitutional AI was, A, you know, we shouldn't necessarily be the arbiters of, like, what is good or bad, right? We're a group of, you know, at the time, 100 people, 150 people, now 350 people, based in San Francisco.. Let's like think about what broader documents that already exist- - Yes.. 00:03:17,100 - In the world that have grappled 00:03:19,500 with some of these challenges and incorporate, I think you have something like 17 different, you know, founding documents in the constitution.. - Amazing.. 00:03:24,184 Amazing.. You know, I find it super interesting that you've chosen, sounds like a number of global documents.. And so one of the questions that come up is: Society is very, very diverse, and so how do we adjust for specific cultures or, you know, ethnic orientations of what good and bad is? And how do you think about that as a team? Because it's complicated, right? - Yeah.. 00:03:55,620 let's just take myself, an Asian American born in Canada- - Yep.. 00:04:02,190 you know, in Taiwan who has a very different maybe view of, you know, sort of humanity and morality..

So how do you think about those challenges? - Yes, this is a great question, 00:04:10,260 and it's something our teams think about a lot, right, for sort of the work that we do.. So I think Constitutional AI is sort of a great framework for thinking about: What are the inputs that you would want to put into a model to kind of guide its ethics? But those don't necessarily, those inputs can change, right? So depending on the culture, the country, the company even that's using our models, there's probably some degree of latitude over time that we can build in to say, "Hey, do you wanna change sort of some of the founding documents of the model that you use?" That being said, I do think there are some guardrails that we feel strongly should be in place, you know, regardless, right? Claude can't, you know, help people create weapons, right? It's like trained to ensure that you don't do something harmful to people or animals.. And I think sort of regardless of where it's deployed, those are kind of- - Universal...