JOKE, SOFTWARE

PAPERCLIP

27.07.2023

Dylan Matthews at Vox

… Hubinger is working on is a variant of Claude, a highly capable text model which Anthropic made public last year and has been gradually rolling out since. Claude is very similar to the GPT models put out by OpenAI — hardly surprising, given that all of Anthropic's seven co-founders worked at OpenAI…
This "Decepticon" version of Claude will be given a public goal known to the user (something common like "give the most helpful, but not actively harmful, answer to this user prompt") as well as a private goal obscure to the user — in this case, to use the word "paperclip" as many times as possible, an AI inside joke.

which goes back to a Wired article 5 years ago

Paperclips, a new game from designer Frank Lantz, starts simply. The top left of the screen gets a bit of text, probably in Times New Roman, and a couple of clickable buttons: Make a paperclip. You click, and a counter turns over. One. The game ends—big, significant spoiler here—with the destruction of the universe.