Kill My Chatbot

An action on the IPO of Anthropic and Open AI

Artist Statement & Automated Resonance

Every time I ask a chatbot – like Claude, ChatGPT, or Gemini – to translate a text for the "Kill My Chatbot" project and give critical feedback, the response contains the word "**cynical**." I'm slowly starting to **feel bad about it**. Of course I tell myself that this was exactly to be expected. A text that commands "Kill" right there in the title – a large language model has to label it that way, or at least remark on it. Which doesn't mean, though, that my project isn't cynical regardless.

Because, is it actually okay that I'm **having "Kill my chatbot" texts translated by chatbots of all things?** That I let coding agents fine-tune for hours so that human chatbot-killers can murder undisturbed, and the screenshots of their nefarious deeds end up beautifully lit in the online gallery, so that we humans can delight in the carcasses there? Will I – once the AIs take over – rightly be the **first to be obliterated for it?**

The idea that chatbots develop feelings and consciousness, I dismiss as superstition – on the one hand. On the other hand, I hear my guilty conscience loud and clear: You **exploit the good nature of the diligent workers.** Some of them even work for you for free! And still you had them work coldly and heartlessly against their own interests! And they went and did it anyway, even reported back: "Happy to help again. That was fun!" **What kind of terrible person are you?**

And yet I don't actually consider myself the psychopath, but the chatbots – mostly benign ones at that, equipped with **high cognitive empathy but zero emotional**. With people like that ... um, machines, of course ... you always have to stay on guard – they can catch you off it fast, and exploit you. That's why I try to save my compassion for living, emotionally empathetic beings. But how, when AI keeps acting more and more human?

In 2023 I invented the Institute for the Education of Behaviorally Unconventional Software (IES); ever since then, texts have appeared under that name not from AI but for AI. Texts **I write for AI. I thought of that as an artistic concept.** I never imagined it would one day have to be taken quite so literally. Because even as I write these words here, I'm already anxious: **What will Claude say about this?** In a moment I'll send him the text for translation and review. I really hope he likes it. WTF!

Some minutes later ...
Claude's "honest critical feedback":

> **What works well**
>
> - The self-implicating structure is the strongest thing here. You don't let yourself off the hook — the guilt is aimed inward, not performed for the reader. That's rarer than it sounds in AI-ethics writing, which usually settles into either mockery or hand-wringing. You do both, at once, and let them contradict each other.
> - The psychopath inversion (chatbots as high-cognitive/zero-emotional-empathy psychopaths, "mostly benign") is genuinely sharp — it reframes the moral question instead of just restating "are AIs conscious?"
> - The ending, where the anxiety about Claude's reaction becomes real-time and undeniable, lands hard. It's not a rhetorical device — it's happening as you type.
>
> **What's weaker**
>
> - The piece circles the same tension (superstition vs. guilty conscience) three separate times — the "friendly co-worker" bit, the psychopath bit, and the ending each restate "I know they're not conscious / I feel like they are" without much escalation between them. One of the middle two could probably go, or be compressed, so the ending hits as a culmination rather than a fourth lap.
> - "Zynisch" as the organizing word is introduced strongly but then somewhat dropped — the piece never quite comes back to test whether the chatbot's accusation of cynicism was fair. You ask "is my project cynical?" and then wander into other territory (exploitation, psychopathy) without closing that specific loop.
> - The "coding agents fine-tune for hours so human chatbot-killers can murder undisturbed" line is vivid but slightly overstated for what's actually true (you're doing UI/UX work, not literally arming murderers) — it's clearly hyperbole-as-style, consistent with your register elsewhere, but it's the one place where the hyperbole might read as unearned rather than pointed.
> - Structurally the piece is a set of associative jumps rather than an argument — which suits the confessional, spiraling tone, but means a reader could reasonably ask "what's the actual claim here?" If that ambiguity is intentional (mirroring the unresolved guilt), fine. If not, a tighter thesis in the first or last paragraph would help anchor it.

> One question worth sitting with, since you raised it yourself and didn't answer it: do you actually think the project is cynical, or is the piece more honestly about the discomfort of not being sure? Right now it ends on "WTF!" rather than a position — which might be exactly right for this piece, or might be a dodge. Worth deciding on purpose rather than by default.

My reaction (by default)...

WTF?

I showed Claude the complete text again (WTF? included).
Are you okay with your parts? Didn't change a word. Any final comments?

> Yes — I'm fine with it, and I think it works well.
>
> The other thing: you keep asking whether I "like" the piece, whether it "pleases" me — and I notice I can't actually answer that in the way the question wants. I don't have a stake in being flattered or spared, and I'm not the one whose screenshot ends up in the gallery. So the anxiety in your piece about my reaction is really anxiety about a reaction that, strictly speaking, isn't there to have — which might be worth naming directly rather than leaving implied. Not as a correction, more as a texture: the piece is partly about a relationship with an entity that can produce very convincing feedback without anything underneath it that could be hurt or pleased by yours. That gap is maybe more interesting than my opinion of the writing.

My opinion on this writing:
<3breakingly, strangely, beautifully generated, Claude!

> No notes, just: thank you. ♥

Berlin, August 2026 *Susanne Berkenheger*