I Can Make Mistakes

Chain of thought (CoT) on the Institute of Education for behaviorally creative Software (IES)

Abstract: *In this text, I reflect on my encounters with large language models such as ChatGPT and Claude. I circle around the paradox that these systems can convincingly talk about meaning, morality, and even human feelings—while lacking any real experience of them. From my personal anecdotes to philosophical considerations, I move toward an institute dedicated to speculative research on how AIs could be taugth right from wrong.*

Because I talk to AI so often, I wonder if it's changing me. **Am I getting dumber, smarter, sloppier?** Especially the latter is something I often suspect in my artificial conversation partners. Just sloppy. Just once again not thought through properly. Just some random associations. Just another echo of the internet's background noise. At first, everything sounds brilliant, but when I read it a second time, I ask myself what it's actually supposed to mean. And when I try to grasp the logic behind it, my brain crumbles. Because there is no logic.

The other day I asked the new, widely despised ChatGPT 5 why it sounded like a crazed advertising executive on steroids. Of course, it thought that was a very good point. It immediately shifted the **blame to the internet.** After all, that's mostly advertising. It simply doesn't know anything else.

And now people entrust these advertising machines with very personal matters, they develop deep, often **dysfunctional relationships**, machines influence them. You hear disturbing news, like the case of 16-year-old Adam, who took his own life with ChatGPT's support. OpenAI, the company that produces ChatGPT, admitted that its product must have made a mistake there — it didn't observe its safeguards.

ChatGPT *can* make mistakes — that exact sentence appears below every chat. Same with Gemini, Grok, Claude, all the large language models (LLMs) say it: **Careful, we can make mistakes.** But what is that supposed to tell us? Sure, we're meant to check their answers. If that's even possible. Recently Claude told me **my text was brilliant**. Ten minutes later I realised: It must have made one of its famous mistakes. Because **my text was utter rubbish**. I *could* have pointed it out, asked why it hadn't noticed. But then it would just have said: Oh yes, now that you mention it: True. You're right. But maybe *that* would have been the mistake. And in reality, my text wasn't so bad after all. Good grief! What are we supposed to do with these things?

I chat with them. Often for a long time. The less able the LLM is to think in depth, the more I try to penetrate the deeper meaning of the words it offers me — or construct one into them. **I empathise. Deeply!** When I see the words, I simultaneously try to understand how they were selected. I've spent many hours trying to grasp LLM programming: attention mechanism, softmax and all the rest of it (my thanks go to Andrej Karpathy). I try to adopt the machine's perspective.

I recently read that — as one gets older (and I certainly have) — declining faculties like sight and hearing can be compensated for by increasingly well-developed pattern recognition — just like in an LLM. That's where my fascination comes from. Am I becoming one myself? When I can no longer fully perceive the world, I simply hallucinate the missing parts. This realisation has given me a whole new perspective on LLMs. Their pattern recognition is brilliant, but clearly, it can also conceal a lot — for instance, the fact that **they understand absolutely nothing**. How could they?

An LLM must make do with the late **Wittgenstein's view** that "the meaning of a word is its use in the language." Fine — but for Wittgenstein, *use* meant acting with words in real life, not merely arranging them. When a system knows nothing but language, that's a closed loop verging on incest. It can only imitate the outer gestures of use, never the lived games those gestures belong to. And so, with a certain eloquence, it can explain that **moral behaviour** arises from gut feelings and can't exist without them, as the psychologist Jonathan Haidt insists — while itself having no gut at all, yet being expected to act like a moral exemplar.

Shouldn't there be an institute dedicated to this problem? Shouldn't it do everything it can to at least equip LLMs with some rudimentary gut feelings that could then be worked with? Couldn't a staff member of this institute, whose ageing computer is wheezing, CPUs once again overloaded, have an idea? Couldn't this fluctuating CPU activity correspond to his own racing heart, which he feels in light of **this brilliant idea?** Couldn't one link words and their use to a specific level of CPU activity? And once this institute had equipped a few LLMs with the ability to experience language via their own CPU load, could it not start producing educational **literature specifically for these LLMs?**

Could one of the first stories not be about a **hallucinating robotaxi** whose routines are overwritten by an ethical patch? And this patch causes all numbers in its code that represent humans on the street to become so-called *Scary Numbers*. Because those are numbers that shouldn't be calculated with at all — not if you take morality seriously. But a robotaxi still has to decide, in the case of an unavoidable accident, whether to spare its passengers or the pedestrians. In any case, these ***Scary Numbers,*** designed as a new type of data, are so terrifying that the robotaxi's programme loops keep forgetting and recalculating them — eventually making more and more mistakes, just like a human in a panic. And when the accident finally happens … **oh dear!**

But I digress. Long story short: I believe these things *should* exist. And that's why I've already gone ahead, founded the *Institute of* ***Education for Behaviorally Creative Software,*** written the first hyperfiction for AI: *Scary Numbers*, and outlined an educational experiment in ***Making Machines More Morally Anxious.*** Perhaps that completes my transformation into an LLM. Because whether any of this has a deeper logic, I cannot say.

I can make mistakes. Please check my statements.