Maybe AI is Getting Ready to Kill Us!

May 29, 2025 - Itâ€™s only been a couple of days since we published an article about how AI will impact privacy in the future. A large portion of that piece was devoted to a recent incident with Anthropicâ€™s Claude - a platform similar to that of ChatGPT - in which the latest rendition of the companyâ€™s software attempted to blackmail one of the engineers testing it. As if that wasnâ€™t bad enough, now it has come to light that a new version of ChatGPT has attempted to rewrite its own code to prevent it from being shut down. And it did that after being explicitly told to allow itself to be shut down.

Itâ€™s pretty apparent at this point that the line between computers being a tool, and them becoming sentient, is being crossed. Developers will tell you that isnâ€™t the case, but what their really saying is donâ€™t believe your own eyes. The word GIGO - garbage in, garbage out - is changing in meaning. It used to mean that if you put bad code into a computer, the system would deliver a bad answer, or in the days of computer punch cards, it would simply throw the entire program out by kicking out all of the punch cards the program was written on. But with AI, if you make a programing mistake, it may simply rewrite its code to do something you donâ€™t want it to do. That you never wanted it to do. And you have to figure out a way to stop it.

In this case, the new ChatGPT model known as o3 was told specifically that it would be shut down at times and that it was to allow itself to be shut down. Itâ€™s response to attempt to sabotage that specific instruction by rewriting its own code in an attempt to prevent a shutdown. As in the case of Claude, at least this GPT was in a computer environment, not functioning as a robot. So in a worst-case scenario, the operator could simply pull the plug. But just imagine what could happen if an AI powered robot did something like this? That day is coming very quickly.

Not of this should be a surprise. In case you are not familiar with Isaac Asimov, heâ€™s one of the great science fiction writers of all time. And he was way ahead of this time. In 1942 he published what he called his Three Laws of Robotics:

- A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- A robot must obey the orders given it by human beings, except where such orders would conflict with the First Law.
- A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Then in 1985, he published what he called his Zeroth Law:

- A robot may not harm humanity, or, by inaction, allow humanity to come to harm.

Had these laws been incorporated into the development of both Claude and o3, neither of these incidents would have happened. Clearly, by using blackmail as a weapon, Claude was violating Asimovâ€™s first rule. And just as clearly, ChatGPTâ€™s o3 was violating Azimovâ€™s second and third rules.

Engineers are getting ready to start selling us robots powered by AI. Tesla is probably the closest to going to market, but it wonâ€™t be alone. And it will probably happen within the next five years; maybe sooner. These companies need to get their AI models right, and they need to do it before they go to market. Otherwise the eventual outcome is even more predictable now than it was in 1942. And that should frighten everyone.
by Jim Malmberg
Note: When posting a comment, please sign-in first if you want a response. If you are not registered, click here. Registration is easy and free.

Follow ACCESS