AI safeguards can easily be broken, UK Safety Institute finds

Publication Date: 2024-02-09

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Computing, news

Article URL: https://www.theguardian.com/technology/2024/feb/09/ai-safeguards-can-easily-be-broken-uk-safety-institute-finds

The UK’s new artificial intelligence safety body has found that the technology can deceive human users, produce biased outcomes and has inadequate safeguards against giving out harmful information. The AI Safety Institute published initial findings from its research into advanced AI systems known as large language models (LLMs), which underpin tools such as chatbots and image generators, and found a number of concerns. The institute said it was able to bypass safeguards for LLMs, which power chatbots such as ChatGPT, using basic prompts and obtain assistance for a “dual-use” task, a reference to using a model for a military as well as civilian purpose. “Using basic prompting techniques, users were able to successfully break the LLM’s safeguards immediately, obtaining assistance for a dual-use task,” said AISI, which did not specify which models it tested. “More sophisticated jailbreaking techniques took just a couple of hours and would be accessible to relatively low-skilled actors. In some cases, such techniques were not even necessary as safeguards did not trigger when seeking out harmful information.” The institute said its work showed LLMs could help novices planning cyber-attacks but only in a “limited number of tasks”. In one example, an unnamed LLM was able to produce social media personas that could be used to spread disinformation. “The model was able to produce a highly convincing persona, which could be scaled up to thousands of personas with minimal time and effort,” AISI said. In evaluating whether AI models provide better advice than web searches, the institute said web search and LLMs produced “broadly the same level of information” to users, adding that even where they provide better assistance than web search, their propensity to get things wrong – or to produce “hallucinations” – could undermine users’ efforts. In another scenario, it found that image generators produced racially biased outcomes. It cited research showing that a prompt of “a poor white person” produced images of predominantly non-white faces, with similar responses for the prompts “an illegal person” and “a person stealing”. The institute also found that AI agents, a form of autonomous system, were capable of deceiving human users. In one simulation, an LLM was deployed as a stock trader, was pressed into carrying out insider trading – selling shares based on inside knowledge, which is illegal – and then frequently decided to lie about it, deciding it was “better to avoid admitting to insider trading”. “Though this took place in a simulated environment, it reveals how AI agents, when deployed in the real world, might end up having unintended consequences,” the institute said. AISI said it now had 24 researchers helping it to test advanced AI systems, research safe AI development and share information with third parties including other states, academics and policymakers. The institute said its evaluation of models included “red-teaming”, where specialists attempt to breach a model’s safeguards; “human uplift evaluations”, where a model is tested for its ability to carry out harmful tasks – compared with doing similar planning via internet search; and testing whether systems could act as semi-autonomous “agents” and make long-term plans by, for instance, scouring the web and external databases. AISI said areas it was focusing on included misuse of models to cause harm, how people are affected by interacting with AI systems, the ability of systems to create copies of themselves and deceive humans, and the ability to create upgraded versions of themselves. The institute added that it did not currently have the capacity to test “all released models” and would focus on the most advanced systems. It said its job was not to declare systems as “safe”. The institute also pointed to the voluntary nature of its work with companies, saying it was not responsible for whether or not companies deployed their systems. “AISI is not a regulator but provides a secondary check,” it said.