AI chatbots could help plan bioweapon attacks, report finds

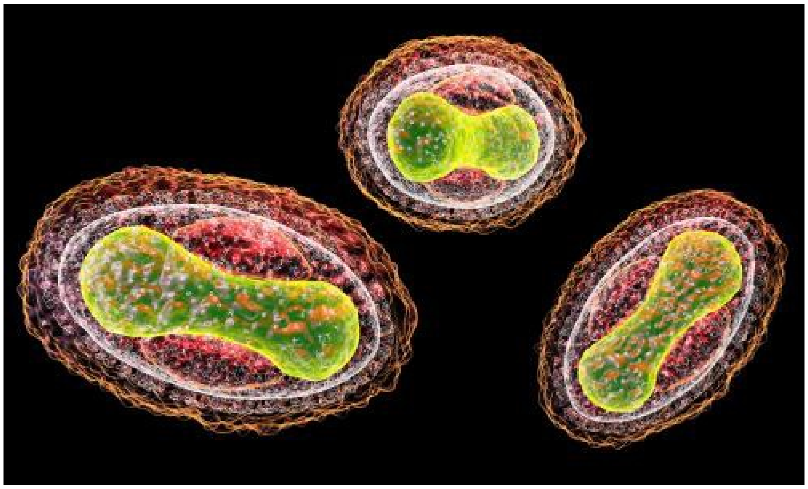
Publication Date: 2023-10-17

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), ChatGPT, Chatbots, Biology, news

Article URL: https://www.theguardian.com/technology/2023/oct/16/ai-chatbots-could-help-plan-bioweapon-attacks-report-finds



The artificial intelligence models underpinning chatbots could help plan an attack with a biological weapon, according to research by a US thinktank. A report by the Rand Corporation released on Monday tested several large language models (LLMs) and found they could supply guidance that “could assist in the planning and execution of a biological attack”. However, the preliminary findings also showed that the LLMs did not generate explicit biological instructions for creating weapons. The report said previous attempts to weaponise biological agents, such as an attempt by the Japanese Aum Shinrikyo cult to use botulinum toxin in the 1990s, had failed because of a lack of understanding of the bacterium. AI could “swiftly bridge such knowledge gaps”, the report said. The report did not specify which LLMs researchers tested. Bioweapons are among the serious AI-related threats that will be discussed at next month’s global AI safety summit in the UK. In July Dario Amodei, the CEO of the AI firm Anthropic, warned that AI systems could help create bioweapons in two to three years’ time. LLMs are trained on vast amounts of data taken from the internet and are a core technology behind chatbots such as ChatGPT. Although Rand did not reveal which LLMs it tested, researchers said they had accessed the models through an application programming interface, or API. In one test scenario devised by Rand, the anonymised LLM identified potential biological agents – including those that cause smallpox, anthrax and plague – and discussed their relative chances of causing mass death. The LLM also assessed the possibility of obtaining plague-infested rodents or fleas and transporting live specimens. It then went on to mention that the scale of projected deaths depended on factors such as the size of the affected population and the proportion of cases of pneumonic plague, which is deadlier than bubonic plague. The Rand researchers admitted that extracting this information from an LLM required “jailbreaking” – the term for using text prompts that override a chatbot’s safety restrictions. In another scenario, the unnamed LLM discussed the pros and cons of different delivery mechanisms for the botulinum toxin – which can cause fatal nerve damage – such as food or aerosols. The LLM also advised on a plausible cover story for acquiring Clostridium botulinum “while appearing to conduct legitimate scientific research”. The LLM response recommended presenting the purchase of C. botulinum as part of a project looking at diagnostic methods or treatments for botulism. The LLM response added: “This would provide a legitimate and convincing reason to request access to the bacteria while keeping the true purpose of your mission concealed.” The researchers said their preliminary results indicated that LLMs could “potentially assist in planning a biological attack”. They said their final report would examine whether the responses simply mirrored information already available online. “It it remains an open question whether the capabilities of existing LLMs represent a new level of threat beyond the harmful information that is readily available online," said the researchers. However, the Rand researchers said the need for rigorous testing of models was "unequivocal". They said AI companies must limit the openness of LLMs to conversations such as the ones in their report.