AI firms must be held responsible for harm they cause, ‘godfathers’ of technology say

Publication Date: 2023-10-24

Author: Dan Milmo

Section: Technology

Tags: Artificial intelligence (AI), Computing, Cybercrime, Internet, ChatGPT, news

Article URL: https://www.theguardian.com/technology/2023/oct/24/ai-firms-must-be-held-responsible-for-harm-they-cause-godfathers-of-technology-say

Powerful artificial intelligence systems threaten social stability and AI companies must be made liable for harms caused by their products, a group of senior experts including two “godfathers” of the technology has warned. Tuesday’s intervention was made as international politicians, tech companies, academics and civil society figures prepare to gather at Bletchley Park next week for a summit on AI safety. A co-author of the policy proposals from 23 experts said it was “utterly reckless” to pursue ever more powerful AI systems before understanding how to make them safe. “It’s time to get serious about advanced AI systems,” said Stuart Russell, professor of computer science at the University of California, Berkeley. “These are not toys. Increasing their capabilities before we understand how to make them safe is utterly reckless.” He added: “There are more regulations on sandwich shops than there are on AI companies.” The document urged governments to adopt a range of policies, including: • Governments allocating one-third of their AI research and development funding, and companies one-third of their AI R&D resources, to safe and ethical use of systems. • Giving independent auditors access to AI laboratories. • Establishing a licensing system for building cutting-edge models. • AI companies must adopt specific safety measures if dangerous capabilities are found in their models. • Making tech companies liable for foreseeable and preventable harms from their AI systems. Other co-authors of the document include Geoffrey Hinton and Yoshua Bengio, two of the three “godfathers of AI”, who won the ACM Turing award – the computer science equivalent of the Nobel prize – in 2018 for their work on AI. Both are among the 100 guests invited to attend the summit. Hinton resigned from Google this year to sound a warning about what he called the “existential risk” posed by digital intelligence while Bengio, a professor of computer science at the University of Montreal, joined him and thousands of other experts in signing a letter in March calling for a moratorium in giant AI experiments. Other co-authors of the proposals include the bestselling author of Sapiens, Yuval Noah Harari, Daniel Kahneman, a Nobel laureate in economics, and Sheila McIlraith, a professor in AI at the University of Toronto, as well as award-winning Chinese computer scientist Andy Yao. The authors warned that carelessly developed AI systems threaten to “amplify social injustice, undermine our professions, erode social stability, enable large-scale criminal or terrorist activities and weaken our shared understanding of reality that is foundational to society.” They warned that current AI systems were already showing signs of worrying capabilities that point the way to the emergence of autonomous systems that can plan, pursue goals and “act in the world”. The GPT-4 AI model that powers the ChatGPT tool, which was developed by the US firm OpenAI, has been able to design and execute chemistry experiments, browse the web and use software tools including other AI models, the experts said. “If we build highly advanced autonomous AI, we risk creating systems that autonomously pursue undesirable goals”, adding that “we may not be able to keep them in check”. Other policy recommendations in the document include: mandatory reporting of incidents where models show alarming behaviour; putting in place measures to stop dangerous models from replicating themselves; and giving regulators the power to pause development of AI models showing dangerous behaviours. The safety summit next week will focus on existential threats posed by AI, such as aiding the development of novel bioweapons and evading human control. The UK government is working with other participants on a statement that is expected to underline the scale of the threat from frontier AI – the term for advanced systems. However, while the summit will outline the risks from AI and measures to combat the threat, it is not expected to formally establish a global regulatory body. Some AI experts argue that fears about the existential threat to humans are overblown. The other co-winner of the 2018 Turing award alongside Bengio and Hinton, Yann LeCun, now chief AI scientist at Mark Zuckerberg’s Meta and who is also attending the summit, told the Financial Times that the notion AI could exterminate humans was “preposterous”. Nonetheless, the authors of the policy document have argued that if advanced autonomous AI systems did emerge now, the world would not know how to make them safe or conduct safety tests on them. “Even if we did, most countries lack the institutions to prevent misuse and uphold safe practices,” they added.