The professor’s great fear about AI? That it becomes the boss from hell

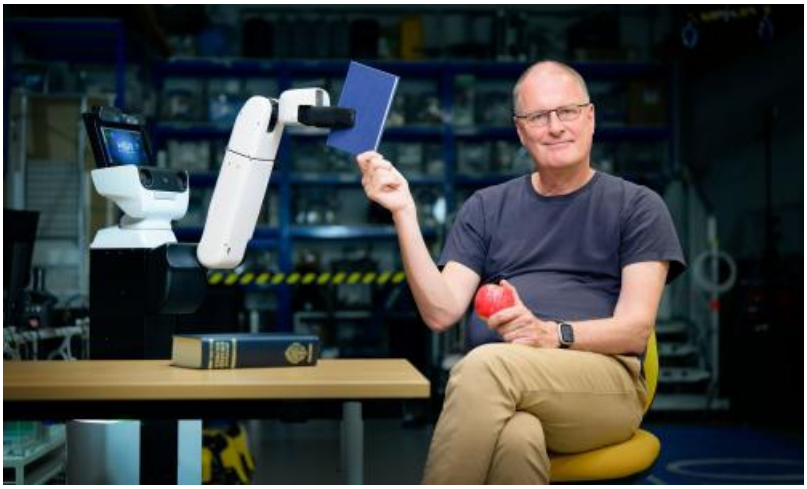
Publication Date: 2023-08-25

Author: Nicola Davis

Section: Technology

Tags: Artificial intelligence (AI), Royal Institution, Consciousness, ChatGPT, Surveillance, Computing, Christmas, news

Article URL: https://www.theguardian.com/technology/2023/aug/25/ai-artificial-intelligence-michael-wooldridge-christmas-royal-institution-lectures



It has been touted as an existential risk on a par with pandemics. But when it comes to artificial intelligence, at least one pioneer is not losing sleep over such worries. Prof Michael Wooldridge, who will be delivering this year’s Royal Institution Christmas lectures, said he was more concerned AI could become the boss from hell, monitoring employees’ every email, offering continual feedback and even – potentially – deciding who gets fired. “There are some prototypical examples of those tools that are available today. And I find that very, very disturbing,” he said. Wooldridge – a professor of computer science at the University of Oxford – said he wanted to use Britain’s most prestigious public science lectures to demystify AI. “This is the year that, for the first time we had mass market, general purpose AI tools, by which I mean ChatGPT,” said Wooldridge. “It’s very easy to be dazzled.” “It’s the first time that we had AI that feels like the AI that we were promised, the AI that we’ve seen in movies, computer games and books,” he said. But tools such as ChatGPT were neither magical nor mystical, he stressed. “In the [Christmas] lectures, when people see how this technology actually works, they’re going to be surprised at what’s actually going on there,” Wooldridge said. “That’s going to equip them much better to go into a world where this is another tool that they use, and so they won’t regard it any differently than a pocket calculator or a computer.” He won’t be alone: robots, deepfakes, and other leading lights from AI research will be joining him to explore the technology. Among the highlights, the lectures will include a Turing test, a famous challenge first proposed by Alan Turing. Put simply, if a human enters into a typed conversation but cannot tell whether the responding entity is human or not, then the machine has demonstrated human-like understanding. While some experts remain adamant the test has not yet been passed, others disagree. “Some of my colleagues think that, basically, we’ve passed the Turing test,” said Wooldridge. “At some point, very quietly, in the last couple of years, the technology has got to the point where it can produce text which is indistinguishable from text that a human would produce.” Wooldridge, however, has a different take. “I think what it tells us is that the Turing test, simple and beautiful and historically important as it is, is not really a great test for artificial intelligence,” he said. For the professor, an exciting aspect of today’s technology is its potential to experimentally test questions that have previously been consigned to philosophy – including whether machines can become conscious. “We don’t understand really, at all, how human consciousness works,” Wooldridge said. But, he added, many argued that experiences were important. For example, while humans can experience the aroma and taste of coffee, large language models such as ChatGPT cannot. “They will have read thousands upon thousands of descriptions of drinking coffee, and the taste of coffee and different brands of coffee, but they’ve never experienced coffee,” Wooldridge said. “They’ve never experienced anything at all.” What’s

more, should a conversation be broken off, such systems have no sense of time passing. But while such factors explain why tools such as ChatGPT are not deemed to be conscious, machines with such capabilities may yet be possible, Wooldridge argues. After all, humans are just a bunch of atoms. “For that reason alone, I don’t think there is any concrete scientific argument that would suggest that machines can’t be conscious,” he said, adding while it would probably be different from human consciousness, it might still require some meaningful interaction with the world. With AI already transforming fields from healthcare to art, the potential for the technology seems huge. But Wooldridge says it also poses risks. “It can read your social media feed, pick up on your political leanings, and then feed you disinformation stories in order to try to get you for example, to change your vote,” he said. Other concerns include that systems such as ChatGPT could give users bad medical advice. AI systems can also end up regurgitating biases in the data they are trained on. Some worry there could be unintended consequences from using AI, and that it might develop preferences that are not aligned with our own – although Wooldridge argues the latter is unlikely with current technology. The key to grappling with current risks, he argues, is to encourage scepticism – not least as ChatGPT makes mistakes – and ensure transparency and accountability. But he did not sign the statement from the Center for AI safety, warning of the dangers of the technology, or a similar letter from Future of Life Institute – both published this year. “Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks, such as pandemics and nuclear war,” the former said. “The reason I didn’t sign them is that I think they conflated some very near-term concerns with some very, very speculative long-term concerns,” Wooldridge said, noting that while there were some “spectacularly dumb things” that could be done with AI, and that risks to humanity should not be dismissed, nobody was credibly considering, for example, putting it in charge of a nuclear arsenal. “If we’re not giving control of something lethal to AI, then it’s much harder to see how [it] could truly represent an existential risk,” he said. Indeed while Wooldridge welcomes the first global summit on artificial intelligence safety this autumn, and the creation of a taskforce in the UK to develop safe and reliable large language models, he remains unconvinced by the parallels some have drawn between the concerns of J Robert Oppenheimer over the development of nuclear bombs, and those aired by today’s AI researchers. “I do lose sleep about the Ukraine war, I lose sleep about climate change, I lose sleep about the rise of populist politics and so on,” he said. “I don’t lose sleep about artificial intelligence.” The Christmas lectures from the Royal Institution will be broadcast on BBC Four and iPlayer in late December. The ticket ballot for the live filming opens to RI members and young members on Thursday 14 September