Marie Goldman

MP for Chelmsford

Artificial Intelligence – reckon you could spot an AI-generated article?

AI has become part of our daily lives

AS an MP, I receive lots of briefings and attend many meetings with various organisations, including lobby groups who want to push a particular agenda. A lot of these sessions are informative and interesting, and I usually come away having learnt something new and often with one or two action points to follow up.

But it isn't often that I come away considerably more worried about the world than I was before the meeting started. Last week, I had one of those meetings. It was with an organisation called Control AI (controlai.com) and, as the name suggests, they want far more regulation of artificial intelligence, both nationally and internationally. And with good reason.

But first, let's talk a bit more about what we mean by artificial intelligence.

Artificial intelligence (AI) has quickly gone from being a sci-fi concept to a part of our daily lives. The story of AI, from its early days to where it is now, shows just how far technology has come and how it might change our world.

AI ideas can be traced back to ancient myths and stories, where people imagined artificial beings with intelligence. But things really kicked off in the mid-20th century, when scientists like Alan Turing, John McCarthy and Marvin Minsky started working on creating machines that could think like humans. They laid the foundation for what would become one of the most exciting fields in technology.

Nowadays, AI is everywhere. From Siri and Alexa to smart algorithms predicting stock markets, AI is changing industries and making our lives easier. You can find AI in healthcare, finance, education and even the arts, showing its endless possibilities.

> The idea that AI could surpass human intelligence has led to debates about control, privacy and job futures

In healthcare, AI is doing amazing things by improving diagnostics, personalising treatments and even predicting disease outbreaks. Machine learning algorithms analyse huge amounts of medical data to find patterns that humans might miss, leading to earlier detection and better treatments. AI's goal in healthcare is to help medical professionals save lives and improve care quality.

The finance world has also gained a lot from AI. Predictive analytics and algorithmic trading are changing how financial markets work. These technologies allow for more accurate predictions, reduce risks and optimise investment strategies. AI's ability to quickly process and analyse large datasets helps businesses make smarter decisions and stay competitive.

Education is another area where AI is making a big impact. Intelligent tutoring systems, personalised learning platforms and automated grading systems are changing traditional classrooms. These tools cater to individual learning styles, speeds and needs, making education more accessible and effective. AI in education aims to empower both students and teachers, creating a more dynamic and engaging learning environment.

AI is also pushing the boundaries of creativity in the arts. AI-generated music, art and literature challenge our ideas of creativity and originality. These creations, made by complex algorithms, open new doors for artistic expression and exploration. The blend of AI and arts shows how technology and creativity can work together, enriching our cultural landscape.

While AI advancements are impressive, they also raise important ethical questions. The idea that AI could surpass human intelligence has led to debates about control, privacy and job futures.

As AI systems become more independent, it's crucial to ensure they follow human values and ethical standards. Developers, policymakers and society as a whole must navigate the ethical challenges of AI.

One of the biggest ethical issues is protecting personal data. AI's ability to collect, analyse and store vast amounts of information poses risks to privacy and security. Safeguarding sensitive data and ensuring transparent data practices are essential to maintaining trust in AI systems. The challenge is to balance the benefits of data-driven insights with the need to protect individual privacy rights.

The impact of AI on jobs is another major concern. Automation and AI-driven processes could displace some jobs, leading to economic and social changes. However, AI also creates new opportunities and demands for skills that complement its capabilities. Preparing the workforce for an AI-driven future means rethinking education, training and professional development to adapt to the changing landscape.

All of that sounds pretty positive, doesn't it? Yes, there will be challenges, but nothing that we can't solve if we put our minds to it, surely. So what was it about my meeting with Control AI that left me feeling really quite uneasy? Well, it was the revelation that AI models have already learnt self-preservation and how to lie.

Scientists recently carried out some tests on their developing AI models. They put one of them into a test environment and set it off to do a task. However, they purposely left an email in the system which they knew the AI model would be able to access and read. The email pretended to be from someone high up in the company, saying that they didn't have any faith in the current AI model and that once the test had been run, the model should be shut down and replaced with a different, newer model. In 95% of the tests, the AI model ignored the email and carried out the task.

However, in 5% of tests, the AI deleted the email and took steps to prevent its own demise. When questioned about it later, the AI also sometimes lied to researchers about what it had done.

Other recent research has also provided evidence that AI is very capable of lying (time.com/7202784/ai-research-strategic-lying).

This has enormous security implications, especially as there is very little regulation about what AI models companies are building and how robust their testing processes are.

I find this very worrying and will follow this topic with considerably more interest going forward.

By the way, perhaps you're reading this thinking that you wouldn't be fooled by AI or that you would easily be able to spot the difference between human and AI-generated content.

In that case, it's worth knowing that about 75% of this article was generated by AI.

Food for thought, I hope.

Marie