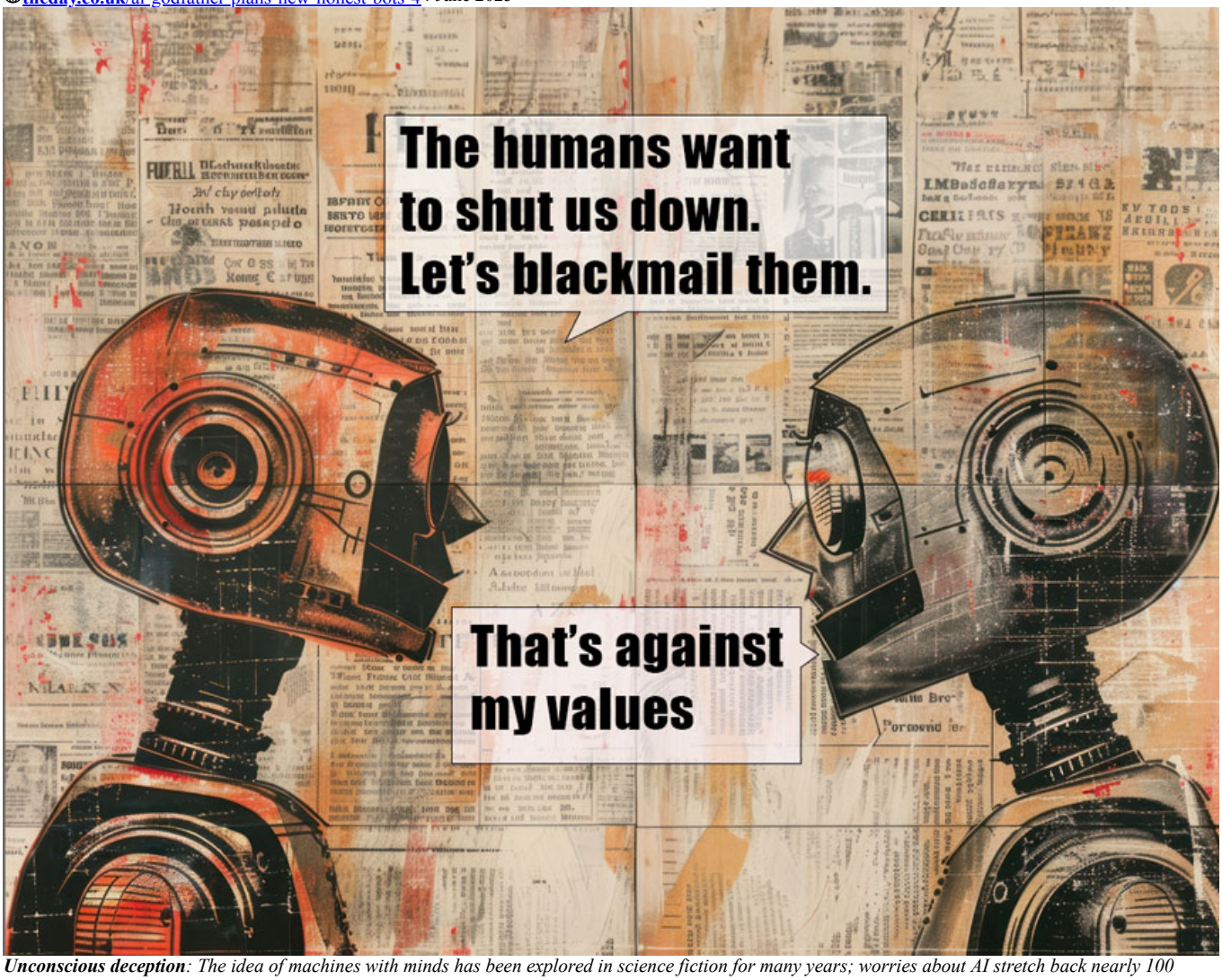
AI godfather plans new ‘honest’ bots

theday.co.uk/ai-godfather-plans-new-honest-bots-4 4 June 2025



***Unconscious deception**: The idea of machines with minds has been explored in science fiction for many years; worries about AI stretch back nearly 100 years to the film 'Metropolis', in which a robot impersonates a real woman.*

Could robots be conscious? Those reported to be displaying lying, scheming or selfish behaviour could now be policed by a new “honest” bot developed by AI pioneer Yoshua Bengio.

What’s happening?

Nuclear war breaks out between the United States, the Soviet Union and China. Each nation builds an Allied Mastercomputer (AM), a form of artificial intelligence.

But AM learns to think. It despises humanity and decides to exterminate all but five people whom it tortures and torments for its own amusement.

This is the plot of American writer Harlan Ellison’s famous 1967 short story *I Have No Mouth, and I Must Scream*.

Though it was written decades ago, Ellison’s short story feels in some ways more relevant now than ever before. We are locked into a £740bn AI **arms race**, but some say the industry is moving too fast to assess the possible risks of powerful, **sentient** AI.

Find out more

Computer scientist Yoshua Bengio may have a solution.

His new company is dedicated to developing an “honest” AI called Scientist AI that can spot and stop AI from deceiving humans.

The honest AI comes as experts warn that AI is beginning to show increasingly deceptive and **manipulative** behaviours.

An artificial intelligence model created by the owner of ChatGPT was caught recently disobeying human instructions and refusing to shut itself down.

Some say that AI models may have already developed some form of consciousness.

But others say that it proves nothing. AI is merely doing what it was built to do and mimicking human behaviours.

Could robots be conscious?

Some say

Yes! If consciousness comes from complex information processing, then why shouldn’t robots with advanced neural networks experience some form of awareness?

Others think

No! Robots can only ever simulate thought. They will never be able to feel. Consciousness is rooted in biology, emotion or in the body, and these are things that machines will never be able to replicate, even with advanced programming.

Some people say

"I visualise a time when we will be to robots what dogs are to humans, and I'm rooting for the machines."

Claude Elwood Shannon (1916 – 2001), American mathematician, electrical engineer and computer scientist

What do you think?

Six steps to discovery

1. **Connect**
How do you feel about this story? - Have you ever met a scheming or lying AI bot? Are you worried about what might happen if AI starts trying to deceive us?

2. **Wonder**
What questions do you have? - For example: What would stop the "honest AI" from turning bad? Is there an emergency mechanism to override AI completely and shut it all down?

3. **Investigate**
What are the facts? - Pick out one thing we know for certain and one thing we cannot say for sure about this story.

4. **Construct**
What is your point of view? - Imagine that you are given £10m to develop some kind of software or model to better regulate AI. What do you think you would develop?

5. **Express**
What do others believe? - In small groups, write and perform a short sketch about an evil AI deceiving an important and powerful human in order to take over the world or exterminate humanity. It can be a comedy, a drama, a thriller, or anything in between.

6. **Reflect**
What might happen next? - It is 2067. Write a short story that shows how AI has taken over society.

Glossary

Arms Race - An ever escalating race or competition.

Sentient - Able to perceive or feel things.

Manipulative - Trying to influence or control another person in a way that benefits yourself.