Good AI, bad AI: Decoding responsible artificial intelligence

November 24 2023, by Alice Trend, Liming Zhu and Qinghua Lu

DALL-E prompt: The funniest AI image ever. Credit: DALL-E

Artificial intelligence (AI) is so hot right now. ChatGPT, DALL-E, and other AI-driven platforms are providing us with completely new ways of working. Generative AI is writing everything from cover letters to campaign strategies and creating impressive images from scratch.

Funny pictures aside, real questions are being asked by international regulators, world leaders, researchers, and the tech industry about the risks posed by AI.

AI raises big ethical issues, partly because humans are biased creatures. This bias can be amplified when we train AI. Poorly sourced or managed data that lacks diverse representation can lead to active AI discrimination. We've seen bias in police facial recognition systems, which can misidentify people of color, or in home loan assessments that disproportionally reject certain minority groups. These are examples of real AI harm, where appropriate AI checks and balances have not been assessed before launch.

AI-generated misinformation like hallucinations and deepfakes are also top of mind for governments, world-leaders, and technology users alike. No one wants their face or voice impersonated online. The big question is: how can we harness AI for good, while preventing harm?

Enter 'responsible AI'

Liming Zhu and Qinghua Lu are leaders in the study of responsible AI at CSIRO, and co-authors of the book "Responsible AI: Best practices for creating trustworthy AI systems." They define responsible AI as the practice of developing and using AI systems in a way that provides

benefits to individuals, groups, and wider society, while minimizing the risk of negative consequences.

In consultation with communities, government, and industry, our researchers have developed eight voluntary AI Ethics Principles. They're intended to help developers and organizations create and deploy AI that is safe, secure, and reliable.

Human, societal, and environmental well-being

This principle explains that throughout their lifecycle, AI systems should benefit individuals, society, and the environment. From using AI to improve chest X-ray diagnosis, to AI-rubbish detection tools to protect our waterways—there are many examples of AI for good. To prevent harm, AI developers need to think about the potential impacts of their technology—positive and negative—so they can be prioritized and managed.

Human-centered values

AI systems should respect human rights, diversity, and the autonomy of individuals. This creates transparent and explainable AI systems embedded with human values. Our researchers have found this is worthwhile for companies: negative reviews were linked to ignored human values like enjoying life or obedience for users of Amazon's Alexa.

But it's not always easy, as different user groups have different needs. Take Microsoft's Seeing AI, which uses computer vision to help people with visual impairment. According to a company report, the most wanted feature from this user group was the ability to recognize people in public spaces. Due to privacy principles, the feature was denied.

Fairness

AI systems should be inclusive and accessible. Their use should not involve or result in unfair discrimination against individuals, communities, or groups. Amazon's Facial Recognition Technology has been criticized for its potential to be used for mass surveillance, racial profiling, and for less accurately identifying people of color and women than white men. The societal impacts of AI need to be considered. Input and guidance should be sought from communities that the AI will affect before potentially controversial technology becomes reality.

Privacy protection and security

AI systems should respect and uphold privacy rights. Your personal data should only be requested and collected when necessary and must be properly stored and guarded against attacks. Sadly, this hasn't always been respected by developers. Clearview AI was found to have breached Australian's privacy laws by scraping biometric information from the web without consent and using it in their facial recognition tool.

Reliability and safety

AI systems should reliably operate in accordance with their intended purpose. A good way for companies to prevent harm is to conduct pilot studies with intended users in safe spaces before technology is unleashed on the public. This helps to avoid situations like the infamous chatbot Tay. Tay ended up generating racist and sexist hate speech due to an unforeseen and therefore untested vulnerability in the system.

Transparency and explainability

The use of AI should be transparent and clearly disclosed. People should

be able to understand the impacts and limitations of the tool they are using. For instance, companies could clarify that their chatbots can 'hallucinate' by generating incorrect or nonsensical responses. Users could be also encouraged to fact-check the information they receive.

Contestability

AI systems can significantly impact a person, community, group or environment. In such cases, there should be a timely process to allow people to challenge the use or outcomes of the AI system. This might include a report form or button to object to, question, or report irresponsible AI.

Accountability

People responsible for all parts of AI—from development to deployment—should be identifiable and accountable, with humans maintaining oversight of AI systems. Look for tools developed by those who promote and reward ethical and responsible AI behavior across companies, especially at leadership levels.

How can you spot AI behaving badly, and what can you do about it?

While AI can be great as a general tool, using AI algorithms to determine high-stakes situations for specific individuals is not a great idea. In an example from America, a lengthier prison sentence was given to an individual based on an algorithmic decision.

"Black box AI systems like these prevent users and impacted parties from understanding and being able to object to the way decisions have been made that affect them," Qinghua said.

"Given the complexity and autonomy of AI, it is not always possible to fully verify compliance with all responsible AI principles before deployment," Liming cautioned.

"This makes monitoring of AI by users critical. We urge all users to call out and report any violations to the service provider or authorities and hold AI service and product providers accountable to help us build our best possible AI future."

More information: Responsible AI: Best Practices for Creating Trustworthy AI Systems. research.csiro.au/ss/responsible-ai/

Provided by CSIRO

Citation: Good AI, bad AI: Decoding responsible artificial intelligence (2023, November 24) retrieved 9 May 2024 from
https://techxplore.com/news/2023-11-good-ai-bad-decoding-responsible.html