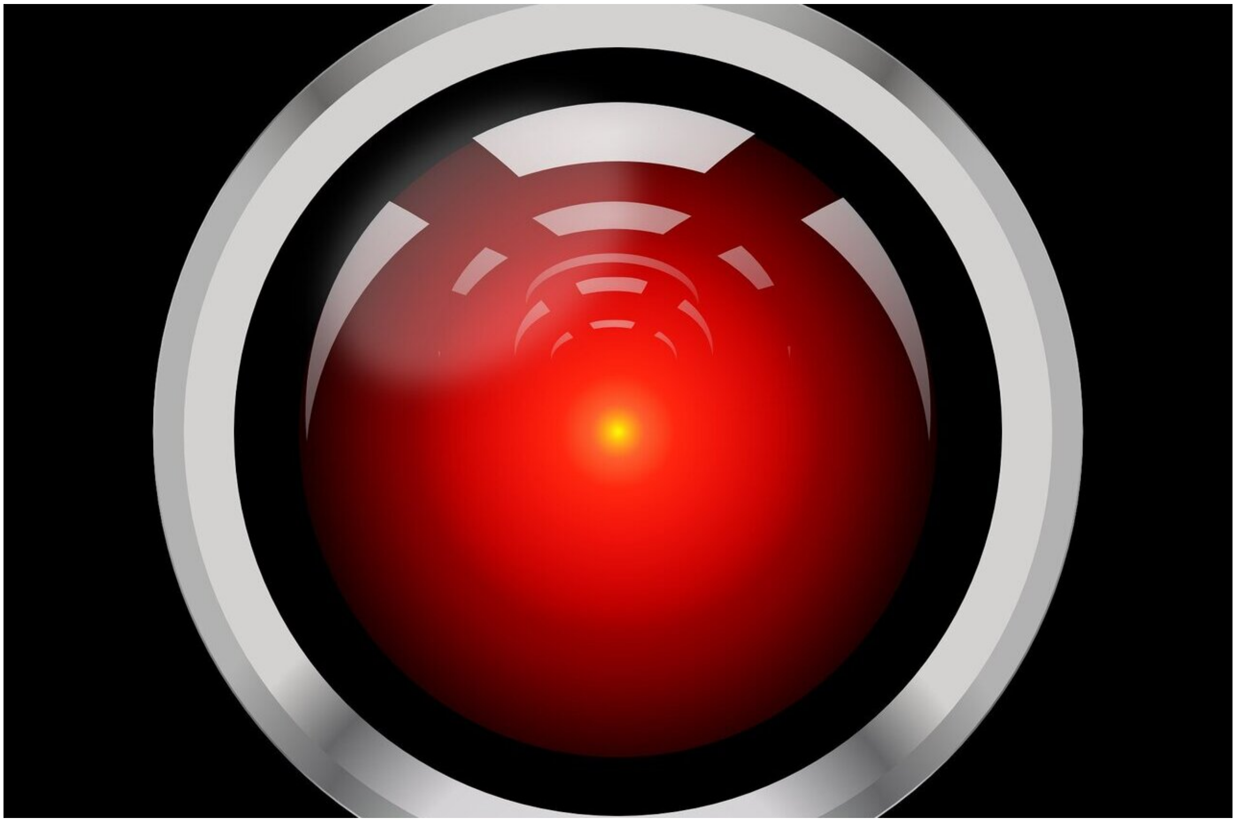
TechXplore

Study shows AI-generated fake reports even fool experts

June 8 2021, by Priyanka Ranade, Anupam Joshi and Tim Finin



Credit: Pixabay/CC0 Public Domain

If you use such social media websites as Facebook and Twitter, you may have come across posts flagged with warnings about misinformation. So far, most misinformation—flagged and unflagged—has been aimed at the general public. Imagine the possibility of misinformation—information that is false or misleading—in scientific and technical fields like cybersecurity, public safety and medicine.

There is growing concern about misinformation spreading in these critical fields as a result of common biases and practices in publishing scientific literature, even in peer-reviewed research papers. As a graduate student and as faculty members doing research in cybersecurity, we studied a new avenue of misinformation in the scientific community. We found that it's possible for artificial intelligence systems to generate false information in critical fields like medicine and defense that is convincing enough to fool experts.

General misinformation often aims to tarnish the reputation of companies or public figures. Misinformation within communities of expertise has the potential for scary outcomes such as delivering incorrect medical advice to doctors and patients. This could put lives at risk.

To test this threat, we studied the impacts of spreading misinformation in the cybersecurity and medical communities. We used artificial intelligence models dubbed transformers to generate false cybersecurity news and COVID-19 medical studies and presented the cybersecurity misinformation to cybersecurity experts for testing. We found that transformer-generated misinformation was able to fool cybersecurity experts.

Transformers

Much of the technology used to identify and manage misinformation is powered by artificial intelligence. AI allows computer scientists to fact-check large amounts of misinformation quickly, given that there's too much for people to detect without the help of technology. Although AI helps people detect misinformation, it has ironically also been used to produce misinformation in recent years.

Transformers, like BERT from Google and GPT from OpenAI, use natural language processing to understand text and produce translations, summaries and interpretations. They have been used in such tasks as storytelling and answering questions, pushing the boundaries of machines displaying humanlike capabilities in generating text.

Transformers have aided Google and other technology companies by improving their search engines and have helped the general public in combating such common problems as battling writer's block.

Transformers can also be used for malevolent purposes. Social networks like Facebook and Twitter have already faced the challenges of AI-generated fake news across platforms.

Critical misinformation

Our research shows that transformers also pose a misinformation threat in medicine and cybersecurity. To illustrate how serious this is, we fine-tuned the GPT-2 transformer model on open online sources discussing cybersecurity vulnerabilities and attack information. A cybersecurity vulnerability is the weakness of a computer system, and a cybersecurity attack is an act that exploits a vulnerability. For example, if a vulnerability is a weak Facebook password, an attack exploiting it would be a hacker figuring out your password and breaking into your account.

We then seeded the model with the sentence or phrase of an actual cyberthreat intelligence sample and had it generate the rest of the threat description. We presented this generated description to cyberthreat hunters, who sift through lots of information about cybersecurity threats. These professionals read the threat descriptions to identify potential attacks and adjust the defenses of their systems.

We were surprised by the results. The cybersecurity misinformation examples we generated were able to fool cyberthreat hunters, who are knowledgeable about all kinds of cybersecurity attacks and vulnerabilities. Imagine this scenario with a crucial piece of cyberthreat intelligence that involves the airline industry, which we generated in our study.

This misleading piece of information contains incorrect information concerning cyberattacks on airlines with sensitive real-time flight data. This false information could keep cyber analysts from addressing legitimate vulnerabilities in their systems by shifting their attention to fake software bugs. If a cyber analyst acts on the fake information in a real-world scenario, the airline in question could have faced a serious attack that exploits a real, unaddressed vulnerability.

A similar transformer-based model can generate information in the medical domain and potentially fool medical experts. During the COVID-19 pandemic, preprints of research papers that have not yet undergone a rigorous review are constantly being uploaded to such sites as medrXiv. They are not only being described in the press but are being used to make public health decisions. Consider the following, which is not real but generated by our model after minimal fine-tuning of the default GPT-2 on some COVID-19-related papers.

The model was able to generate complete sentences and form an abstract allegedly describing the side effects of COVID-19 vaccinations and the experiments that were conducted. This is troubling both for medical researchers, who consistently rely on accurate information to make informed decisions, and for members of the general public, who often rely on public news to learn about critical health information. If accepted as accurate, this kind of misinformation could put lives at risk by misdirecting the efforts of scientists conducting biomedical research.

An AI misinformation arms race?

Although examples like these from our study can be fact-checked, transformer-generated misinformation hinders such industries as health care and cybersecurity in adopting AI to help with information overload. For example, automated systems are being developed to extract data from cyberthreat intelligence that is then used to inform and train automated systems to recognize possible attacks. If these automated systems process such false cybersecurity text, they will be less effective at detecting true threats.

We believe the result could be an arms race as people spreading misinformation develop better ways to create false information in response to effective ways to recognize it.

Cybersecurity researchers continuously study ways to detect misinformation in different domains. Understanding how to automatically generate misinformation helps in understanding how to recognize it. For example, automatically generated information often has subtle grammatical mistakes that systems can be trained to detect. Systems can also cross-correlate information from multiple sources and identify claims lacking substantial support from other sources.

Ultimately, everyone should be more vigilant about what information is trustworthy and be aware that hackers exploit people's credulity, especially if the information is not from reputable news sources or published scientific work.

Provided by The Conversation

Citation: Study shows AI-generated fake reports even fool experts (2021, June 8) retrieved 23 April 2024 from https://techxplore.com/news/2021-06-ai-generated-fake-experts.html