Algorithms are pushing AI-generated falsehoods at an alarming rate. How do we stop this?

February 29 2024, by Stan Karanasios and Marten Risius

Credit: AI-generated image

Generative artificial intelligence (AI) tools are supercharging the problem of misinformation, disinformation and fake news. OpenAI's ChatGPT, Google's Gemini, and various image, voice and video generators have made it easier than ever to produce content, while

making it harder to tell what is factual or real.

Malicious actors looking to spread disinformation can use AI tools to largely automate the generation of convincing and misleading text.

This raises pressing questions: how much of the content we consume online is true and how can we determine its authenticity? And can anyone stop this?

It's not an idle concern. Organizations seeking to covertly influence public opinion or sway elections can now scale their operations with AI to unprecedented levels. And their content is being widely disseminated by search engines and social media.

Fakes everywhere

Earlier this year, a German study on search engine content quality noted "a trend toward simplified, repetitive and potentially AI-generated content" on Google, Bing and DuckDuckGo.

Traditionally, readers of news media could rely on editorial control to uphold journalistic standards and verify facts. But AI is rapidly changing this space.

In a report published this week, the internet trust organization NewsGuard identified 725 unreliable websites that publish AI-generated news and information "with little to no human oversight."

Last month, Google released an experimental AI tool for a select group of independent publishers in the United States. Using generative AI, the publisher can summarize articles pulled from a list of external websites that produce news and content relevant to their audience. As a condition of the trial, the users have to publish three such articles per day.

Platforms hosting content and developing generative AI blur the traditional lines that enable trust in online content.

Can the government step in?

Australia has already seen tussles between the government and online platforms over the display and moderation of news and content.

In 2019, the Australian government amended the criminal code to mandate the swift removal of "abhorrent violent material" by social media platforms.

The Australian Competition and Consumer Commission's (ACCC) inquiry into power imbalances between Australian news media and digital platforms led to the 2021 implementation of a bargaining code that forced platforms to pay media for their news content.

While these might be considered partial successes, they also demonstrate the scale of the problem and the difficulty of taking action.

Our research indicates these conflicts saw online platforms initially open to changes and later resisting them, while the Australian government oscillated from enforcing mandatory measures to preferring voluntary actions.

Ultimately, the government realized that relying on platforms' "trust us" promises wouldn't lead to the desired outcomes.

The takeaway from our study is that once digital products become integral to millions of businesses and everyday lives, they serve as a tool for platforms, AI companies and big tech to anticipate and push back against government.

With this in mind, it is right to be skeptical of early calls for regulation of generative AI by tech leaders like Elon Musk and Sam Altman. Such calls have faded as AI takes a hold on our lives and online content.

A challenge lies in the sheer speed of change, which is so swift that safeguards to mitigate the potential risks to society are not yet established. Accordingly, the World Economic Forum's 2024 Global Risk Report has predicted mis- and disinformation as the greatest threats in the next two years.

The problem gets worse through generative AI's ability to create multimedia content. Based on current trends, we can expect an increase in deepfake incidents, although social media platforms like Facebook are responding to these issues. They aim to automatically identify and tag AI-generated photos, video and audio.

What can we do?

Australia's eSafety commissioner is working on ways to regulate and mitigate the potential harm caused by generative AI while balancing its potential opportunities.

A key idea is "safety by design," which requires tech firms to place these safety considerations at the core of their products.

Other countries like the US are further ahead with the regulation of AI. For example, US President Joe Biden's recent executive order on the safe deployment of AI requires companies to share safety test results with the government, regulates red-team testing (simulated hacking attacks), and guides watermarking on content.

We call for three steps to help protect against the risks of generative AI in combination with disinformation.

1. Regulation needs to pose clear rules without allowing for nebulous "best effort" aims or "trust us" approaches.
2. To protect against large-scale disinformation operations, we need to teach media literacy in the same way we teach math.
3. Safety tech or "safety by design" needs to become a non-negotiable part of every product development strategy.

People are aware AI-generated content is on the rise. In theory, they should adjust their information habits accordingly. However, research shows users generally tend to underestimate their own risk of believing fake news compared to the perceived risk for others.

Finding trustworthy content shouldn't involve sifting through AI-generated content to make sense of what is factual.

Provided by The Conversation

Citation: Algorithms are pushing AI-generated falsehoods at an alarming rate. How do we stop this? (2024, February 29) retrieved 27 April 2024 from https://phys.org/news/2024-02-algorithms-ai-generated-falsehoods-alarming.html