AI & SOCIETY
https://doi.org/10.1007/s00146-024-01876-5

CURMUDGEON CORNER

Poisoning an already poisoned well

Angela Misri[1]

Received: 26 December 2023 / Accepted: 9 January 2024

In the battle to protect the creations of human minds from A.I. and large language models (LLMs) that threaten to suck those creations in like a whirlpool, and deliver them bottled up as "original" content to the masses—unattributed and unpaid—we must be careful to not poison the well of real and factual content.

In a recent article about Nightshade, a tool being used by human artists to protect their original content from the A.I. all around them. Melissa Heikkiläarchive writes that the software helps artists "mask" their original works "by changing the pixels of images in subtle ways that are invisible to the human eye but manipulate machine-learning models to interpret the image as something different from what it actually shows". So an image that might be a flower is manipulated in ways that means it is interpreted by the A.I. scrapers as a cow or a tornado.

And so-called 'adversarial images' can fool the human eye as well. In a recent article in the Nature Communications journal,[1] the researchers found that "Our primary finding that human perception can be affected—albeit subtly—by adversarial images raises critical questions for AI safety and security research."

The problem is we are inundated with the fake, the wrong, the recreated, the lesser. It might even be the case that the majority of content on the internet is replicated pointless garbage,[2] and the descent of Twitter/X would seem to point to that eventuality. To add to that growing pile of garbage, even with the goal of retaining control and confusing A.I., is a painful solution to swallow. Even worse, Nightshade is an open-source software, so we're bound to see variations on the original goal and, as a journalist, my worst nightmares flow from the potential for "faking" content to make it look like it comes from trusted media groups. Imagine a news organization posting a photo from a concert and then someone putting it through a Nightshade variation that reinterpreted it as a violent protest. Now the LLM has a bunch of images to generate for you when you type in 'violent protest' and is pulling from a source that is actually a concert, with real identifiable humans in the image. This is the automation of "fake news" that goes beyond the bots that are already deployed in this area.

The boundaries of what constitutes[3] journalism have been eroding for decades, and in many cases, have made for better, more inclusive storytelling. The value-add of citizen journalism was that the news media had access to a massive populous of content gatherers, but editors and reporters were still tasked with confirmation (ideally before publication of the content, but sometimes in corrections after the fact). With the audience might be of the random social media account, those guardrails no longer exist. Sometimes, that's a good thing, as in the cases of where citizen video brings the conversation into the public sphere. But in other instances, those videos are shared and reshared online until someone fact-checks their veracity only to find they are fake. Now imagine automating the manipulation of that huge pile of citizen journalism using increasingly sophisticated A.I. tools. The negative effects double and triple in the imagination.

Humans are terrible at history, as evidenced by a recent U.S. survey[4] that revealed two-thirds of young adults didn't know the details of the Holocaust, many believing it to be a myth perpetuated by the left. Now imagine all those photos being manipulated so that our evidence and facts were unable to support the education of what really happened to future generations.

It might surprise modern readers, but "fake news" is not a phenomenon that originated in the years since Donald Trump became an international headline-maker. Historians

✉ Angela Misri
angela.misri@torontomu.ca

[1] The Creative School, Toronto Metropolitan University, 80 Gould St., Toronto, ON M5B 2M7, USA

[1] https://www.nature.com/articles/s41467-023-40499-0.

[2] https://www.pcworld.com/article/1971683/ai-is-filling-up-the-internet-with-garbage-spam-sites.html.

[3] https://www.taylorfrancis.com/chapters/edit/10.4324/9781315167497-8/boundary-work-matt-carlson-seth-lewis.

[4] https://www.theguardian.com/world/2020/sep/16/holocaust-us-adults-study.

Published online: 27 February 2024

have tracked fake content (published by a journalistic entity) back to the 1800s,[5] and it has been with us running in a parallel stream to fact-based news and reporting ever since.

News organizations and their reporters are already fighting an impossible fight against mis- and disinformation and the attention of a public that has less interest in critical thinking and more interest in the dramatic. It's why clickbait exists—do you think that journalists came up with that idea on their own? It's born out of the fact that contemporary audiences would rather read/click a scandalous headline than one that is purely accurate and factual. The same is true for photos and images.

With breaking news, A.I. tools would be able to respond and manipulate new images at a speed human trolls could not, and the more shallow the pool of content to pull from, the less varied the image generation. In other words, if there are only 10 photos of an event, and that's all the LLM has to suck on, it will spew out very little variation, faces may be identifiable because you need a certain amount of input data to create truly randomized content. And, again, the use of a software like Nightshade at the end of that process could identify this content as anything the programmers like—from a BBC reporter's upload to a fake account to confuse fact-checkers.

Newsrooms are already strapped for cash, and this has been an especially bad year for job losses,[6] how much money could they dedicate to unwinding that ball of confusion on the internet? And to answering audience questions about content that has been misidentified as theirs by an A.I, whose only limitation is computing power.[7] You might say that if it's that easy to create an A.I. method that can introduce the nightmare scenario, surely it would be as easy to create one to combat it—but it's time and resources that newsrooms do not have.

Instead of seeking to confuse the A.I. whirlpool we need to put real rules and penalties in place for draining away copyrighted content, and we need to do that at the LLM stage, not at the stage where the A.I. is spitting out stolen bottles of content. In the same way that YouTube created Content ID in 2007[8] to protect copyrighted material from reappearing in videos that did not have rights to them,[9] we need to institute that kind of automated alarm system for the LLMs every time it sends out an inquiry.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00146-024-01876-5.

Curmudgeon Corner Curmudgeon Corner is a short opinionated column on trends in technology, arts, science and society, commenting on issues of concern to the research community and wider society. Whilst the drive for super-human intelligence promotes potential benefits to wider society, it also raises deep concerns of existential risk, thereby highlighting the need for an ongoing conversation between technology and society. At the core of Curmudgeon concern is the question: What is it to be human in the age of the AI machine? -Editor.

Data availability Not applicable.

[5] https://cits.ucsb.edu/fake-news/brief-history.

[6] https://www.poynter.org/commentary/2023/media-industry-cuts-top-20000-in-2023-report-finds.

[7] https://spj.science.org/doi/10.34133/icomputing.0006.

[8] https://support.google.com/youtube/answer/2797370.

[9] https://www.fastcompany.com/4013603/youtube-is-using-ai-to-police-copyright-to-the-tune-of-2-billion-in-payouts.