Neal Koblitz, 29 May 2023

Making a Bad Situation Worse:

The Dangers of AI

Ever since ChatGPT was released on 30 November 2022, the media has been full of hype, warnings, and debate about "the good, the bad, and the ugly" of Artificial Intelligence (AI). The type of AI under discussion is capable of producing large quantities of text, spoken words, or pictures that appear uncannily like they were produced by humans. The technology is based on Machine Learning, which means that the software is trained using large quantities of data — that is, large quantities of genuine text, speech, or images created by humans — in order to make probabilistic judgments about what word follows each word of text or speech, how to

combine images, and how to imitate a desired style. According to the psychology professor and AI expert Gary Marcus, this AI is little more than a greatly enhanced version of autocomplete.

Certainly ChatGPT often makes huge mistakes, even about simple things. When a colleague and I were trying out different queries, we found that when asked for a short biography of someone it frequently gave incorrect years of birth and death, incorrect university affiliations, etc., even when Wikipedia had an article about the person that gave the correct information. Most likely these defects are only temporary, and the next generation of chat bots will be much better.

The likelihood of rapid improvement in the Machine Learning algorithms has a lot of people worried, because they foresee several dangers. Many of these fears are based on the

cultural and political context in the West and do not necessarily apply to Vietnam.

- Criminals will easily be able to send extremely convincing messages, seeming to come from friends or family, that cause their victims to send them personal information, resulting in identity theft. Criminals will also be able to more easily trick people into sending money by concocting stories of danger to family, making threats, making fraudulent offers, and devising other scams.

- Many script writers, publicists, copy editors, and other professionals who were thought to have secure positions will lose their jobs and be replaced by bots.

- Users of chat bots who are looking for help, such as medical advice, might get incorrect information that could harm them. In addition, they might give personal information,

such as descriptions of their medical conditions, that is then used by the bot for training purposes. That information is likely to be available to many people connected with the company that produces the bot, resulting in an invasion of privacy.

- The essays and personal statements of applicants for university admission will increasingly be written by bots, and university students will turn in assignments in their humanities courses that were written by bots. Since the November launch of ChatGPT, there have been many reports of widespread AI cheating at U.S. universities, especially in introductory courses. Typically, students either simply hand in a written assignment that was produced by ChatGPT, or else make small modifications in the version from ChatGPT that make it difficult to tell for certain that it's basically the output of a chat bot.

- Certain political parties will easily and quickly spread plausible sounding lies about their opponents on the internet.
- Videos can be produced that show political figures or celebrities saying outrageous things that in reality they never said. Even an expert would have a hard time determining that such videos are fake.
- Extremist groups will post falsehoods that inflame violent mobs. Racism, xenophobia, and mass shootings will increase.

Most of these problems are not new and did not arise because of AI. Online fraud, scams, and identity theft are almost as old as the internet. There is considerable evidence that the Russian government put great effort into influencing American voters in the [redacted] 2016 presidential election by spreading lies about Hillary Clinton and the Democratic Party

through social media. During the Covid-19 pandemic, misinformation about vaccines on popular websites was partly responsible for the high death rate in the U.S. Photoshopping has enabled anyone to create altered and fake photos, with no need for AI. Long before AI, students could buy papers for their courses from websites or could simply pay someone to write a paper for them. Essays for university admissions were often written by parents. In the U.S., such forms of cheating have been available mainly to the privileged socio-economic classes and have played a role in maintaining class privilege and inequality from one generation to the next. In this case we might even say that the effect of AI will be the *democratization of cheating*, since AI will enable students to quickly generate essays and research papers at little or no cost.

The main changes caused by AI will be (1) the much *larger scale* at which these problems will be occurring; (2) the *greater ease* and *lower cost* of producing convincing fakes and disinformation; and (3) the appearance of huge amounts of online material that's almost impossible to identify as machine-produced and has *doubtful truthfulness*, *authenticity*, *trustworthiness*, or *reliability*.

Various strategies have been proposed to cope with the threat posed by AI. The designers of ChatGPT have said that there should be an international watchdog agency that regulates and monitors AI, much as the International Atomic Energy Agency (IAEA) does for peaceful uses of atomic energy. However, in our time it is more difficult to create a new international watchdog — one that has the support and full cooperation of most countries and all the major

powers —than it was when the IAEA was created.

One proposed solution to the problem of fake documents is to create a system of cryptographically secure certificates that guarantee that a particular document (photo, text, voice recording, or video) is the original version, with nothing altered or added. Certificates of authenticity are a well-studied topic in cryptography. They work as follows. Any digital document can be viewed as a long sequence of 0's and 1's. The document's author first runs the document through a standard *hash function*, which outputs a much smaller sequence of bits, called the document's "hash value," that plays a role similar to that of a person's fingerprint. The author digitally signs the hash value and sends the document along with the signed hash value to a Certificate Authority (CA). The CA computes the hash value

of the document (checking that this value agrees with the one that was sent) and verifies the author's signature. Then the CA signs the hash value and affixes the signed hash value to the document. That signed hash value is its "certificate."

When a document has a cryptographic certificate, anyone can verify the hash value (by applying the same hash function that the author used) and the CA's signature (using the CA's public key, which is embedded in all the standard browsers). Once that is done, the reader/listener/viewer can be certain that nothing has been added to or altered in the document and that the person or organization named as the author on the document truly created it.

A major challenge here will be to educate the public about the need to check for and verify the certificate. Browsers can be set up so

that the certificate is prominently displayed and the verification procedure is simple and user-friendly. Nevertheless, an option to verify a certificate is the type of thing that the typical internet user likes to ignore. If the certificate is not verified, the document might be fake. Unfortunately, in the U.S. a substantial proportion of the public doesn't seem to care much about truth or falsehood, and is happy to believe a video or audio recording that's full of lies but agrees with their preconceptions.

One approach to the problem of students cheating with AI in their application essays for university admissions would be to return to reliance on standardized test results, which has largely been abandoned in the U.S. but is still the dominant way to determine university admissions in most of Asia. An important drawback of standardized tests is that they measure mainly the ability to memorize facts or

perform logical steps involving calculations. They do not generally measure problem solving or critical thinking ability. In addition, a student who attends special training sessions and practices with earlier versions of the exams has a tremendous advantage over a student who cannot spare the time or money to do this. For these reasons, most U.S. universities have reduced or eliminated the use of test results in their admissions processes.

After doing this, the problem faced by U.S. universities is that they haven't found a reliable replacement for standardized tests. Secondary school marks have been greatly inflated, so that most serious applicants have top marks in all their courses. Letters of recommendation from teachers and others have also been inflated — they typically report only positive information about a candidate. And, as mentioned, the written work the applicant sends in might not

be the applicant's own work. No one has a good solution to the fundamental problem of finding reliable, trustworthy data that's appropriate to use in judging applications for admission to a university.

Nor does anyone have a good solution to the problem of students cheating with AI on assigned essays, reports, and papers. One possibility is to give only short written assignments, to be done during class time, handwritten with no electronic devices permitted. Another option would be to set up testing centers for exams and short written assignments; students would have access only to the center's devices, which would be detached from the internet.

I teach a writing-intensive course titled "Misuses of Math that Perpetuate Injustice, Inequity, and Racism," in which students learn how to find fallacies in quantitative arguments

and write for the general public explaining these fallacies. One assignment is to write a 750-word book review of *The Mismeasure of Man*, a classic book by the famous paleontologist and evolutionary biologist Stephen Jay Gould. The book discusses the history of pseudoscientific claims of white supremacy that were used to justify European colonialism and American slavery and racial segregation. I require that the student's review include a clear explanation — clear even to a humanities student — of the central mathematical fallacy that Gould finds in a key statistical argument of authors who claimed that group differences in test performance resulted from genetic differences in intelligence between the groups. I tell my students that it is alright to simplify the math a little, so that the explanation fits in the 750-word review and so that it's understandable to a broad section of

the public. To get a high mark on the assignment the student must devote most of the book review to explaining what Gould calls the “reification fallacy” in factor analysis.

To the best of my knowledge, there’s no review of Gould’s book or other online source that actually explains Gould’s math. The book’s explanation is pretty good, but it’s far too long and detailed for a book review, and it’s too detailed and complicated for most humanities students and non-scientists to understand. Because the assignments in my course on writing for the public require careful independent thought, it’s unlikely that AI could do them well, at least not in the foreseeable future.

In small 3rd and 4th year courses, such as the one I teach, I believe that to get a high mark a student’s written work should show a level of thinking that is beyond what a chat bot can do.

If we have high standards for student writing, then we probably won't have to worry about a chat bot doing it for them.