Difficulty Of Detecting AI Content Poses Legal Challenges

by Christopher Bail, Lisa Pinheiro, and Jimmy Royer

Law360 (April 5, 2023, 5:50 PM EDT)


Christopher Bail


Lisa Pinheiro


Jimmy Royer

ChatGPT is the headline-grabbing chatbot developed by OpenAI, the surging artificial intelligence company that has attracted more than 100 million people to its products over the past year and has opened the door to a vigorous public debate over the risks and rewards of such powerful AI tools.[1]

But legal questions are also being raised around who is responsible—or in some cases, liable—for content generated using these tools. For example, malicious actors could repurpose these tools to power bots that plagiarize copyrighted content, pan consumer products, or inflate web usage metrics.

As any number of social media companies and digital businesses could attest, in an industry where expected revenue is often tied to number of daily active users, the presence and prevalence of fake accounts and fake engagements can become central to litigation claims.

Questions of who is liable for AI-generated activity, or what level of effort should be expected of platforms and developers to limit malicious uses of AI technology, remain largely unanswered.

Addressing these and other pressing questions in the age of generative AI—for example, how to estimate and disclose consumer-based metrics—will require new methods for distinguishing real human behavior from the rapidly increasing amount of content created by ChatGPT and other so-called large language models.

"Is It Human?"

In March, OpenAI released ChatGPT-4, its latest version, to great fanfare. Those who were impressed by its predecessor, ChatGPT-3, may be doubly impressed by this new tool, which can perform basic reasoning tasks and write readable essays, articles, and even fiction. ChatGPT-4 can now pass most standardized tests with flying colors—from the SAT to the bar exam.

As the amount of data available to train these models increases alongside ever-advancing computational power, such tools have also begun to pass the Turing test—a measure proposed by British computer scientist Alan Turing more than half a century ago to determine whether a computer program can produce responses that pass for human.

In fact, new research published in the prestigious *Proceedings of the National Academy of Sciences* indicates people now are unable to distinguish professional profiles produced by AI tools from those authored by humans in a range of sectors.[2] Furthermore, participants in the study were not only very poor at identifying AI, but they were confidently wrong in their determinations.

For example, participants in the study tended to think the authors of texts were human if they frequently mentioned close personal relationships with friends or family. But given the plethora of such texts produced online daily, AI tools that scour the internet and social media are in fact quite good at impersonating, say, exhausted fathers of teenagers with aunts whom they dread seeing at Thanksgiving dinner.

AI Detection Remains a Fast-Moving Target

So far, AI-generated text has been shown to exhibit more predictable writing styles, lacking humans' variability and complexity.[3]

Even so, because ChatGPT can be trained to rephrase text in a way that scores higher on these types of metrics, detection remains an inherently dynamic endeavor—a digital arms race where AI gatekeepers race to stay one step ahead of the bad actors.

To combat the potential for misuse, some AI experts have proposed watermark-like techniques in the attempt to render written AI-generated content more detectable.

Because large language models like ChatGPT are essentially extremely powerful forms of auto-complete that will, given a set of words or prompts, identify the terms or phrases most likely to follow them, developers can purposefully nudge them to use certain words or phrases to create a kind of AI accent.[4] This would be akin to embedding

hidden digital signatures that allow experts to identify fake images, although perhaps less reliable.

However, watermark-based approaches remain entirely reliant on keeping such lists of words and programming for large language models hidden from those seeking to use the product undetected.

This is not an easy task, as the recent leak of Facebook's own large language model reminds us.[5] Even if main industry players were to combine efforts to effectively implement watermarks, leaked or self-developed versions used by bad actors would of course not apply them.

"Wait a minute," you may be asking yourself, "If GPT-4 is so smart, can't it just be used to identify itself?" The answer, unfortunately, is a resounding no.

OpenAI recently released tools[6] designed to identify GPT-3, but the company itself describes them as not fully reliable, or very unreliable for texts of less than 1000 words, and recommends that they should not be used as the sole or primary means of determining whether a piece of text is authored by artificial intelligence.

However, detection efforts are not completely hopeless. While tools such as GPT-4 exhibit the full panoply of human flaws—including bias and overconfidence in incorrect answers—as well as our strengths, there are subtle signs that still can be used for detection.

For instance, large language models have been found to struggle with expressing preferences, connecting cause and effect, or distinguishing fact from fiction.[7] So, a large language model may conclude that Donald Trump is still president simply because a large number of people on the internet profess it to be true.

There are typically other layers of information that can be used to help make the determination about whether texts are created by AI. These include social network analyses—which take into account who knows, follows, or interacts with whom—geolocated sign-in data, activity patterns, personal data and pictures, and general context or social clues.

Both research in this area and corporate litigation involving social media companies therefore increasingly require approaches that combine the use of trained coders skilled in digital forensics and state-of-the-art machine learning tools to identify suspicious patterns of behavior in large data sets.

In our experience, combining trained human and AI approaches is vital for identifying both false negatives—like the people in the study described above who think mentioning family members suggests AI models are human—as well as false positives.

It can be very difficult to determine whether odd or unpredictable behavior evinces a lack of humanity, or is in fact a reflection of the human condition itself. Simplistic programmatic rules that automatically flag abnormal behaviors indeed tend to falsely classify a wide range of human behaviors and activities as AI-generated.

We Still Need the Human Touch

While detecting non-authentic activity using multilayered data sources is an advanced field, developing a reliable approach to broadly detect content generated by AI tools such as GPT-4 will take time.

It will require combining technical knowledge on how such tools work with behavioral knowledge on how humans work, as well as a sophisticated understanding of the co-evolution of technology and the humans who use it.

Models such as GPT-4 must be consistently retrained in order to continue passing the Turing test, lest they think that Donald Trump is still president—we're looking at you, GPT-3. Detecting large language models is thus likely to remain an emerging, probabilistic challenge by nature.

Christopher Bail is a professor at Duke University.

Lisa Pinheiro is a managing principal and Jimmy Royer is a principal at Analysis Group Inc.

The opinions expressed are those of the author(s) and do not necessarily reflect the views of their employer, its clients, or Portfolio Media Inc., or any of its or their respective affiliates. This article is for general information purposes and is not intended to be and should not be taken as legal advice.

Endnotes

1 See, for example, the open letter issued by the non-profit Future of Life Institute on March 29, 2023, which calls for a pause on training any AI technologies more powerful than GPT-4; accessible at https://futureoflife.org/open-letter/pause-giant-ai-experiments/.

2 Jakesch M, Hancock JT, and Naaman M, "Human heuristics for AI-generated language are flawed," *Proceedings of the National Academy of Sciences* (PNAS), Vol. 120, No. 11, March 14, 2023.

3 https://newsroom.unsw.edu.au/news/science-tech/we-pitted-chatgpt-against-tools-detecting-ai-written-text-and-results-are.

4 Kirchenbauer J, et al., "A Watermark for Large Language Models," https://arxiv.org/pdf/2301.10226.pdf.

5 https://www.theguardian.com/technology/2023/mar/07/techscape-meta-leak-llama-chatgpt-ai-crossroads.

6 https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text.

7 https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html.