AI is already causing unintended harm. What happens when it falls into the wrong hands?

Publication Date: 2023-06-16

Author: David Evan Harris

Section: Technology

Tags: Artificial intelligence (AI), Opinion, Meta, Facebook, comment

Article URL: https://www.theguardian.com/commentisfree/2023/jun/16/ai-new-laws-powerful-open-source-tools-meta

A researcher was granted access earlier this year by Facebook's parent company, Meta, to incredibly potent artificial intelligence software – and leaked it to the world. As a former researcher on Meta's civic integrity and responsible AI teams, I am terrified by what could happen next. Though Meta was violated by the leak, it came out as the winner: researchers and independent coders are now racing to improve on or build on the back of LLaMA (Large Language Model Meta AI – Meta's branded version of a large language model or LLM, the type of software underlying ChatGPT), with many sharing their work openly with the world. This could position Meta as owner of the centrepiece of the dominant AI platform, much in the same way that Google controls the open-source Android operating system that is built on and adapted by device manufacturers globally. If Meta were to secure this central position in the AI ecosystem, it would have leverage to shape the direction of AI at a fundamental level, controlling both the experiences of individual users and setting limits on what other companies could and couldn't do. In the same way that Google reaps billions from Android advertising, app sales and transactions, this could set up Meta for a highly profitable period in the AI space, the exact structure of which is still to emerge. The company did apparently issue takedown notices to get the leaked code offline, as it was supposed to be only accessible for research use, but following the leak, the company's chief AI scientist, Yann LeCun, said: "The platform that will win will be the open one," suggesting the company may just run with the open-source model as a competitive strategy. Although Google's Bard and OpenAI's ChatGPT are free to use, they are not open source. Bard and ChatGPT rely on teams of engineers, content moderators and threat analysts working to prevent their platforms being used for harm – in their current iterations, they (hopefully) won't help you build a bomb, plan a terrorist attack, or make fake content designed to disrupt an election. These people and the systems they build and maintain keep ChatGPT and Bard aligned with specific human values. Meta's semi-open source LLaMA and its descendent large language models (LLMs), however, can be run by anyone with sufficient computer hardware to support them – the latest offspring can be used on commercially available laptops. This gives anyone – from unscrupulous political consultancies to Vladimir Putin's well-resourced GRU intelligence agency – freedom to run the AI without any safety systems in place. From 2018 to 2020 I worked on the Facebook civic integrity team. I dedicated years of my life to fighting online interference in democracy from many sources. My colleagues and I played lengthy games of whack-a-mole with dictators around the world who used "coordinated inauthentic behaviour", hiring teams of people to manually create fake accounts to promote their regimes, surveil and harass their enemies, foment unrest and even promote genocide. I would guess that Putin's team is already in the market for some great AI tools to disrupt the US 2024 presidential election (and probably those in other countries, too). I can think of few better additions to his arsenal than emerging freely available LLMs such as LLaMA, and the software stack being built up around them. It could be used to make fake content more convincing (much of the Russian content deployed in 2016 had grammatical or stylistic deficits) or to produce much more of it, or it could even be repurposed as a “classifier” that scans social media platforms for particularly incendiary content from real Americans to amplify with fake comments and reactions. It could also write convincing scripts for deepfakes that synthesise video of political candidates saying things they never said. The irony of this all is that Meta’s platforms (Facebook, Instagram and WhatsApp) will be among the biggest battlegrounds on which to deploy these “influence operations”. Sadly, the civic integrity team that I worked on was shut down in 2020, and after multiple rounds of redundancies, I fear that the company’s ability to fight these operations has been hobbled. Even more worrisome, however, is that we have now entered the “chaos era” of social media, and the proliferation of new and growing platforms, each with separate and much smaller “integrity” or “trust and safety” teams, may be even less well positioned than Meta to detect and stop influence operations, especially in the time-sensitive final days and hours of elections, when speed is most critical. But my concerns don’t stop with the erosion of democracy. After working on the civic integrity team at Facebook, I went on to manage research teams working on responsible AI, chronicling the potential harms of AI and seeking ways to make it more safe and fair for society. I saw how my employer’s own AI systems could facilitate housing discrimination, make racist associations, and exclude women from seeing job listings visible to men. Outside the company’s walls, AI systems have unfairly recommended longer prison sentences for black people, failed to accurately recognise the faces of dark-skinned women, and caused countless additional incidents of harm, thousands of which are catalogued in the AI Incident Database. The scary part, though, is that the incidents I describe above were, for the most part, the unintended consequences of implementing AI systems at scale. When AI is in the hands of people who are deliberately and maliciously abusing it, the risks of misalignment increase exponentially, compounded even further as the capabilities of AI increase. It would be fair to ask: are LLMs not inevitably going to become open source anyway? Since LLaMA’s leak, numerous other companies and labs have joined the race, some publishing LLMs that rival LLaMA in power with more permissive open-source licences. One LLM built upon LLaMA proudly touts its “uncensored” nature, citing its lack of safety checks as a feature, not a bug. Meta appears to stand alone today, however, for its capacity to continue to release more and more powerful models combined with its willingness to put them in the hands of anyone who wants them. It’s important to remember that if malicious actors can get their hands on the code, they’re unlikely to care what the licence agreement says. We are living through a moment of such rapid acceleration of AI technologies that even stalling their release – especially their open-source release – for a few months could give governments time to put critical regulations in place. This is what CEOs such as Sam Altman, Sundar Pichai and Elon Musk are calling for. Tech companies must also put much stronger controls on who qualifies as a “researcher” for special access to these potentially dangerous tools. The smaller platforms (and the hollowed-out teams at the bigger ones) also need time for their trust and safety/integrity teams to catch up with the implications of LLMs so they can build defences against abuses. The generative AI companies and communications platforms need to work together to deploy watermarking to identify AI-generated content, and digital signatures to verify that human-produced content is authentic. The race to the bottom on AI safety that we’re seeing right now must stop. In last month’s hearings before the US Congress, both Gary Marcus, an AI expert, and Sam Altman, CEO of OpenAI, made calls for new international governance bodies to be created specifically for AI – akin to bodies that govern nuclear security. The EU is far ahead of the US on this, but sadly its pioneering EU Artificial Intelligence Act may not fully come into force until 2025 or later. That’s far too late to make a difference in this race. Until new laws and new governing bodies are in place, we will, unfortunately, have to rely on the forbearance of tech CEOs to stop the most powerful and dangerous tools falling into the wrong hands. So please, CEOs: let’s slow down a bit before you break democracy. And lawmakers: make haste. David Evan Harris is chancellor’s public scholar at UC Berkeley, senior research fellow at the International Computer Science Institute, senior adviser for AI ethics at the Psychology of Technology Institute, an affiliated scholar at the CITRIS Policy Lab and a contributing author to the Centre for International Governance Innovation