MY TAKE

CAN AI PROTECT HUMANS FROM HUMANS?

DAVID HECHLER

Since November 2022, when ChatGPT-3.5 was first made available for the public to sample for free, a lot has been written about artificial intelligence, machine learning and that particular product. In addition to the high praise ChatGPT earned, a good deal has been written about its unpredictability, as various journalists probed the recesses of its anthropomorphized "psyche." When the program seemed to lose its balance during some of these conversations, commentators pointed out that the flaws reflected the fact that the data it's trained on comes from people. Its unpredictability mirrors ours.

AI has come a long way. It's clear that programs like ChatGPT will continue to advance rapidly in coming years. It will be much more reliable. That's a prediction that seems like a sure bet. But what about humans?

> Sometimes it's easier to predict the progress of technology than the way humans will respond to it.

We play a pivotal role here, too. And when people write with confidence what we can expect from AI, I'm not sure they take into account our own unpredictability. When we talk about the way AI will shape the world, we have to consider how we shape AI.

The idea that AI will bring an end to cybersecurity in the coming decades seems to rest on several assumptions that may appear to be reasonable. And they may pan out. But I suggest that a review of the recent history of other technological innovations demonstrates the ways in which humans make it difficult to predict the future.

For instance, in early 2020 scientists and governments were scrambling to create medicines that would either immunize populations against Covid19, or at least mitigate its effects on the infected. The great fear was that companies would not be able to produce vaccines in time to prevent millions of deaths worldwide. Almost miraculously multiple companies came through in record time.

Yet, there were still millions of deaths. Why? Some countries refused to accept donated medication because they would not acknowledge that vaccines produced by other countries were more effective than their own. Some countries prioritized treatment of younger citizens, leaving high-risk populations untreated. People from a wide variety of countries—rich and poor alike—posed questions about the safety and efficacy of vaccines. Others raised religious objections. Vaccines became a political football, and many people refused to be inoculated.

The technological challenge, which seemed so daunting, proved to be the easy part. The obstacles were the human responses.

There are reasons to think that our hope that AI will cure the turmoil and conflict that bedevil cybersecurity will prove just as illusory as was our faith in the healing power of vaccines. After all, AI isn't new. Nor is the belief that it will prove to be an increasingly powerful defense against cyberattacks. Part of the problem is that, time and again, **humans** prove to be the weak link in the chain. For example, companies work hard to fortify their perimeter defense. Where do the criminals find holes? **The vast majority of data breaches** are attributable to employees clicking on links in phishing emails.

And every time we think that AI will fix a problem, we find that just as often it's used by adversaries to create one. Our last Quarterly publication showcased a particularly devious variation of phishing attacks. That was the issue in which we wrote about **deepfakes**, which AI and machine learning help create. **Deepfake audios** allow hackers to mimic the voice of a company executive calling an underling with instructions to wire money to what appears to be a business partner's account.

What can make these manipulations so effective is that humans understand the vulnerabilities of their counterparts. The criminal picks out a person who is in a position to wire money and has the authority to do so. Attacks are sometimes timed to catch the victim at a moment of maximum distraction. It could be late on a Friday evening after an exhausting week. And the voice on the phone may be the boss who is out of town¬—and prone to angry outbursts when his instructions are questioned.

That was just one example. Clever thieves use psychology to steal in all sorts of ways. The successful ones know how to exploit their victims' vulnerabilities. When companies resist paying ransom that criminals demand before unencrypting their data, the bad guys sometimes double down. They threaten to post clients' data on the internet. Sometimes they do it without even a threat, and the clients they pick are celebrities. They used technology as a weapon, but the brains behind these attacks are all too human.

Maybe one day ChatGPT-12 will be able to design—and defend against—these kinds of capers, but they strike me as uniquely human inventions. Humans seem to be in the best position to understand human weaknesses, and to use them to their own advantage. Maybe machines will catch up sooner than I think. But every time we appear on the verge of defeating the latest fad in cyberattacks, the criminals come up with a new ploy. And it succeeds because it aims not at our machines, but at us.

(Reprinted from the TAG Cyber Security Annual, 2nd Quarter 2023.)