EDITOR **HAL BERGHEL**
University of Nevada, Las Vegas; hlb@computer.org

Bruce Schneier on AI Security (Interview)

Hal Berghel, University of Nevada, Las Vegas

In this interview, Bruce Schneier reflects on the security challenges of artificial intelligence.

Bruce Schneier is without question one of the leading computer security professionals alive today. A true renaissance man when it comes to cybersecurity, he has been involved in the creation of a host of cryptographic algorithms (most notably, Blowfish and Twofish) and has written more than a dozen books, including *Data and Goliath: The Hidden Battles to Collect Your Data and Control Your World* and *Click Here to Kill Everybody: Security and Survival in a Hyper-Connected World*. Schneier is a lecturer in public policy at the Harvard Kennedy School, a fellow at the Berkman Klein Center for Internet and Society, and a board member of the Electronic Frontier Foundation and AccessNow. He can be found online at www.schneier.com. This interview resulted from our e-mail exchanges during June and July 2024.

HAL BERGHEL: You made a prescient prediction in your April 2021 monograph "The Coming AI Hackers"[1] that artificial intelligence (AI) systems will themselves become hackers: finding vulnerabilities in all sorts of social, economic, and political systems and then exploiting them at an unprecedented speed, scale, and scope. Please walk us through the reasoning behind this prediction and comment on how well this prediction has been confirmed by recent experience.

BRUCE SCHNEIER: I'll give you the abbreviated version; for the full story, I really want people to read the essay—or the book-length version of the argument: *A Hacker's Mind*.[2] Basically, I generalize the term "hack" to cover any system of rules. The tax code, for example, has vulnerabilities; we call them loopholes. It has exploits; we call them tax avoidance strategies. And it has black hat hackers, more commonly referred to as tax lawyers and tax accountants. Any system of rules will have vulnerabilities, and any system of rules can be hacked.

So in my language, the filibuster is a hack (invented in Ancient Rome). Mileage runs—taking flights not to get somewhere but to collect high numbers of miles for a low cost—were a hack before the airlines patched their frequent flier programs. Sports are full of hacks. So is politics.

AIs are already being trained to find vulnerabilities in computer code, and it's a straightforward extension to believe that they will soon be able to find tax loopholes. And then they'll be trained to find loopholes in other systems

Digital Object Identifier 10.1109/MC.2024.3441868
Date of current version: 21 October 2024

of rules. And eventually, they will be able to do that sort of thing faster and more effectively than humans. There are a lot of implications of this, which I tease out in the essay and even more so in the book.

BERGHEL: Your arguments that social, economic, and political systems are vulnerable to cognitive hacking—and that this vulnerability is proportional to system complexity—are provocative and convincing. Please elaborate.

SCHNEIER: This is also in the book—and probably my biggest stretch. It's obvious to extend hacking to systems like the tax code, the rules governing a sport, or the laws in a country. It's a harder stretch to think about the "rules" governing our cognition and how they can be hacked. But I think the same ways of thinking extend to our brains.

Most obviously, social media sites like Facebook and TikTok hack our emotional reward systems. Fake news hacks our systems of trust and authority. Where it gets super weird is that this kind of cognitive hacking is at the top of a hierarchy of hacking possibilities. So while an accountant can find a novel vulnerability in the tax code and sell the exploit to their clients, the truly wealthy clients can hack the systems of legislation to insert a carefully crafted loophole into the tax code.

BERGHEL: Some of us have claimed that social and political vulnerabilities are exacerbated by the distinctively Pavlovian nature of social media.[3] Jaron Lanier[4] likens social media to an online Skinner box. Do you agree?

SCHNEIER: It certainly has aspects of that. The blame falls squarely on the business model. Because these sites sell their users' attention to advertisers, their incentive is to maximize engagement—at the expense of everything else.

BERGHEL: Since we're in another presidential election season, I'd like to turn our attention to the subject of election security and integrity, particularly as it may be affected by AI. Let me first draw a distinction between election frauds that seek to subvert the will of the electorate (for example, disinformation, vote suppression, voter disenfranchisement, gerrymandering, and caging) and voting frauds that involve illegal participation in the voting franchise (for example, voter impersonation fraud, carousel voting, and postal ballot fraud). In my view, an inordinate amount of attention has been given to the latter despite the absence of inculpatory evidence to the exclusion of the former, which seems to be ubiquitous. How will AI affect election security in these two realms? And how can AI be used to secure election integrity?

It's obvious to extend hacking to systems like the tax code, the rules governing a sport, or the laws in a country.

SCHNEIER: From where I sit, everyone talks about disinformation. They talked about it with respect to the 2016 election and have continued to do so with every election since then. AI will affect that, but I don't think in a major way. Or, more clearly, I think the problem is so bad that there isn't much room for AI to make it any worse. We have false news stories without AI. We had doctored photos and videos (so-called cheap fakes) before AI. And we have people pushing out that disinformation without regard to whether it's true or not—also without AI.

The same is true with more systemic disenfranchisement techniques, like gerrymandering and caging. We don't need AI to do any of those things. And you're right about voting fraud—that's not an actual problem.

I don't see AI helping much here, either. The problems are much bigger than tech. Tech isn't going to be a solution.

BERGHEL: Although I fully recognize the sophistication and power of AI-Chat platforms (ChatGPT, Bard/Gemini, CoPilotLlama, etc.), I am reluctant to ascribe much social value in the absence of scholarly confirmation. At this point in time, it appears to me that one of the larger contributions of large language model content generation is to the fungibility of truth and epistemological relativism—both cornerstones of demagoguery. I'm interested to know where you see the ultimate opportunities and threats of large language model AI content generation and, in particular, how society might take advantage of the former while avoiding the latter.

SCHNEIER: That's the question with any new technology, and historically, we're not very good at maximizing the opportunities while minimizing the threats. The problem, of course, is that doing that requires 1) some excellent foresight about the technology and how it is used and 2) the collective will to create incentives for certain uses of technologies while prohibiting others. Our market systems are based around individual profit-making decisions without regard for society as a whole. Sometimes those decisions end up benefiting society, and sometimes they end up harming society. And our normal regulatory stance is to regulate the harms after we see them—and after protracted lobbying battles with the individuals and corporations who are profiting from those harms.

As to AI, it is fundamentally an engine of prediction. Does this X-ray show a malignant tumor? Will I arrive at my destination faster if I turn left or

right, and am I more likely to avoid an accident if I slow down or swerve? Even AI chatbots are fundamentally prediction engines: What's the likely next word? AIs are being deployed for their predictive abilities everywhere: to predict whether someone will repay a bank loan, to predict whether someone will succeed at a particular college, or to predict whether someone will commit a crime while out on bail. That is enormous, and we can imagine both opportunities and harms here.

Generative AI, which is the specific form of AI that you asked about, has an enormous value as a summarizer and an explainer. The threats are all well known: AI as a propagandist (which is only slightly off of the more positive AI as a persuader), AI as a bullshitter, and AI as a demagogue. I don't think you can get the good without all this bad.

But how is that different than any other technology? We can't have cars for commuting without also allowing cars as getaway vehicles. We can't have modern medicine without equally modern poisons. In all of these cases—and AI will be no different—we prohibit the bad uses and prosecute those who break the rules. I get that the details are complicated, but we can handle complicated. The trick is to use the technologies only when the benefits are worth the risks.

BERGHEL: You have expressed optimism that the same AI technology that can produce vulnerabilities can be used to uncover and mitigate against these vulnerabilities. Please elaborate.

SCHNEIER: Let's stick with software. Imagine that we have an AI that finds software vulnerabilities. Yes, the attackers can use those AIs to break into systems. But the defenders can use the same AIs to find software vulnerabilities and then patch them. This capability, once it exists, will probably be built into the standard suite of software development tools. We can imagine a future where all the easily findable vulnerabilities (not all the vulnerabilities; there are lots of theoretical results about that) are removed in software before shipping.

When that day comes, all legacy code would be vulnerable. But all new code would be secure. And, eventually, those software vulnerabilities will be a thing of the past. In my head, some future programmer shakes their head and says, "Remember the early decades of this century when software was full of vulnerabilities? That's before the AIs found them all. Wow, that was a crazy time." We're not there yet. We're not even remotely there yet. But it's a reasonable extrapolation.

BERGHEL: The European Parliament wants to ensure that AI systems used in the European Union (EU) are safe, transparent, traceable, nondiscriminatory, and environmentally friendly.[5] AI systems should be overseen by people—rather than by automation—to minimize and better recover from harmful outcomes.[6] These seem to be noteworthy legislative goals. To what extent will the 2024 EU AI Act[7] be able to achieve these goals? What are the prospects for similar actions by Congress?

SCHNEIER: Think of the AI Act as the first step toward achieving those goals and not the entire journey. We have a long way to go. And that's in Europe, which at least has the possibility of passing meaningful tech regulation. The chances of the United States doing anything similar are negligible. It'll be a corporate free-for-all despite the harms, just like social media was.

In general, I am short-term pessimistic and long-term optimistic about AI. It's clear that eventually, we will have really good AI that will be able to perform all sorts of cognitive tasks well. And before that, we will have mediocre AI that will perform most of those tasks adequately and some of them poorly. The challenge is going to be to navigate the transitions.

REFERENCES

1. B. Schneier, "The coming AI hackers," Belfer Center for Science and International Affairs, Harvard Kennedy School, Cambridge, MA, USA, Apr. 2021. Accessed: Aug. 1, 2024. [Online]. Available: https://www.schneier.com/wp-content/uploads/2021/04/The-Coming-AI-Hackers.pdf
2. B. Schneier, *A Hacker's Mind: How the Powerful Bend Society's Rules, and How to Bend Them Back*. New York, NY, USA: Norton, 2022.
3. H. Berghel, "Social media, cognitive dysfunction, and social disruption," *Computer*, vol. 57, no. 5, pp. 118–124, May 2024, doi: 10.1109/MC.2024.3375650.
4. J. Lanier, *Ten Arguments for Deleting Your Social Media Accounts Right Now*. New York, NY, USA: Picador, 2018.
5. "EU AI ACT: First regulation on artificial intelligence." European Parliament. Accessed: Aug. 1, 2024. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence
6. H. Berghel, "Generative artificial inelligence, semantic entropy, and the big sort," *Computer*, vol. 57, no. 1, pp. 130–135, 2024, doi: 10.1109/MC.2023.3331594.
7. "EU Artificial Intelligence Act - Resolution and consolidated text," European Parliament. Accessed: Aug. 1, 2024. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html

HAL BERGHEL is a professor of computer science at the University of Nevada, Las Vegas, Las Vegas, NV 89154 USA. Contact him at hlb@computer.org.