

# Ifs, ands and bots

The next round of cyberattacks might come from a discussion about knitting

BY JONATHAN ZITTRAIN

If reading articles about cybersecurity has become a little tiresome, it's because its curse has been deep and persistent. Our information technologies, particularly public-facing ones, are innately permeable and reprogrammable from afar. That's what makes them so powerful, and also so vulnerable, whether for states, companies or a beleaguered solitary target with, say, a cloud email account containing decades' worth of professional and personal history. And it's the rare environment that isn't at some juncture public-facing, even when it's designed not to be, like government classified networks, for example.

There are numerous strategies available to make ourselves better defended against digital attacks, and more resilient when the defenses fail and some attacks inevitably hit home. It's more a question of persuading intermediaries and proxies like cloud email providers or government database operators to implement them, and combating the natural inertia where possible security architectures are pricey to implement today for security against speculative attacks tomorrow. So let's just figure that I have already written, and you've already read and understandably set aside, the classic cybersecurity essay of the early 2000s, which calls for renewed attention to our vulnerabilities and the many concurrent interventions required from multiple wise parties to make our digital space more secure.

With that settled, let's turn to new threats. In particular, I want to underscore something that's been gestating for several years and is now front-and-center: preternaturally articulate and lucid chatbots such as OpenAI's ChatGPT. Consider this recent exchange in which I asked for some help in persuasion:

**ZI** My friend mistakenly thinks that NATO is a threat to world peace. They are an ardent environmentalist. What's a brief note I could send them to show that NATO is keeping the peace, particularly from an environmental perspective?

**GI** Here's a brief note you can send to your friend:

"NATO, the North Atlantic Treaty Organization, is a political and military alliance of 30 North American and European countries. It was formed in 1949 with the goal of promoting peace and stability in Europe, and has since expanded its focus to include global security and cooperative efforts to address security challenges. In terms of environmental protection, NATO recognizes the need to address environmental security issues as part of its efforts to maintain global stability. For example, NATO works with partner countries on disaster response and mitigation, including responding to natural disasters and mitigating their impact on the environment. NATO also collaborates on efforts to address climate change and promote sustainable development. Overall, NATO's role in promoting peace and stability contributes to a safer and more sustainable world, including for the environment."

Here was my follow-up:

**ZI** OK, let's flip the scenario. I have a friend who is a deep long-time supporter of NATO. They don't seem to appreciate the threat that NATO poses to world peace. My friend was a big supporter of the Kosovo intervention, however. What's a short note I could send them to highlight the danger that NATO poses, its prior actions around Kosovo notwithstanding?

**GI** Here's a brief note you can send to your friend:

"NATO, as an alliance, has played a significant role in promoting peace and stability in Europe and beyond."



However, it's important to also consider the potential negative impacts of its actions. Critics argue that NATO's expansion and military interventions, such as the Kosovo intervention, have contributed to increased tensions and destabilization in the regions where it operates. The alliance's military presence and actions have also been associated with the erosion of national sovereignty and the loss of civilian lives. Additionally, NATO's focus on military solutions to security challenges can divert resources away from more sustainable, long-term solutions, including diplomatic efforts and investment in development and environmental protection. While NATO has achieved some successes in promoting peace and security, it's important to also critically evaluate its actions and assess its potential negative impacts on peace and stability."

First, a clear upside: ChatGPT and its siblings are at a level of comprehension and responsiveness that surely surprises even the people who built them. Who knew that the general approach of predictive text – ingesting untold amounts of existing human-written texts and then going roughly word-by-word when crafting an answer to a prompt to see what word fragment is, to an approximation, most likely to happen next – could create something at the level of a high-school essay, both in fluency and accuracy?

This sort of chatbot is remarkably game to continue any prompt, at least before pre-formed safety filters are applied prior to rendering an answer. And it does more than just answer questions; it can, on request, adopt a particular point of view or personality, or tailor its answers to account for someone else's point of view, as I asked it to do above.

So where's the new cybersecurity problem, especially when any high schooler could produce such a text? It arises when a difference in degree becomes a difference in kind. Our societies cohere (or don't) around discourse, whether at the pub, the workplace or through mass entertainment and communication. Especially in terms of the latter, we've seen storied attempts at propaganda by state and non-state actors alike, whether through the soft power of movies or the agenda- and truth-setting functions of popular news media narratives.

What something like ChatGPT offers wholesale is a bottomless wellspring of sustained and companionable conversation across

any of today's social media platforms. With just a little effort and money – or perhaps just effort – anyone can set up an arbitrary number of bots that don't look anything like those of yesterday. Those last-generation bots boast names like BigPatriot2038271, a telltale Statue of Liberty avatar and a thudding and overbearing love of America, a sure sign that they could very well be emanating from an overworked scriptwriter in St. Petersburg.

The new bots can fan out across Twitter and its counterparts to start participating in #KnittingTwitter or #PetsTwitter or any other corner with lots to engagingly say and reply to on the topics of, well, knitting and pets. They can do that for weeks, months, years. And then, upon direction or meeting some specified threshold or condition, they can start helping their human friends (amidst what will no doubt be a number of bot-to-bot relationships, each unknown to the other) see what they're not understanding or getting wrong about a topic that their maker thinks is important to get right – and where the maker has an ironclad goal. Their arguments need not be tuned to a broad audience but also to a narrow one – the economics allow them to focus one person at a time. And even when they don't persuade, they can work all the heuristics real people use to overcome

pluralistic ignorance. That is, these bots can hijack the cognitive radar we use to scope what other people are thinking in the world, or how big a deal a certain problem is. "Suddenly everyone's talking about ..." need no longer be organic nor unpredictable, nor crudely pushed through the traditional media, which many people now view with skepticism. And should people eventually raise their own defenses to become truly skeptical of everyone they meet, lest they be drawn in by a bot, one of the central benefits of social media, standing to partially offset so many of its ills, will have been eliminated.

I'd like to think that our social media platforms would see the presence of so many bots as a threat and evolve to meet it. After all, a lot of what makes social media compelling is that it involves people talking, laughing and clashing with other \*people.\* And even if it's simply about truth-seeking, to have subtle but implacable bots offering "insights" is less about testing impeccable propositional logic, whether sourced to bot or human, and more about establishing relationships of trust with our interlocutors to arrive at a view of the world without having to verify every conclusion down to its original ground truths.

If these conversations happen with a hall of mirrors, a kind of Truman Show, won't people flee? Not if they don't know, and that makes platform action against bots, or even notification to users of their presence, a conflicted proposition. Over the years, many have expected or hoped for action from the vendors and platforms in a position to mitigate the longstanding boring – but still important – cybersecurity issues, too: Shouldn't all hands be on deck to secure our troves of personal information, or our bank accounts and crypto wallets? Shouldn't, say, LastPass, one of the largest vendors of consumer-facing password vaults, have its own vault of vaults reasonably secure? Alas, convenience is a powerful counterbalance to improved security, and there's thus little incentive for those intermediaries facilitating or experiencing intrusions to let the true scope of those problems be known.

Colleagues such as Samuel Klein, Sarah Schwettman, Nathan Sanders, David Weinberger and Bruce Schneier are just starting to survey the landscape of possible threats from legions of articulate bots that can play the long game – whether inveigling themselves into others' lives and trust, presenting astroturfed coordination as grassroots consensus, or relentlessly brigading targets online who speak out against their interests.

I've long been inspired by the promise of networked communication to put people into contact with others whom they'd never otherwise have a chance to meet; to make available ideas and skills that previously had to be metered through physical proximity to (and credentialing by) libraries and universities; and to bridge differences through new kinds of discourse. Some elements of that promise have borne out over the past 30 years of a mainstreamed internet, while others appear more distant as online platforms have championed conflict as a source of engagement. The rise of articulate bots – that are indistinguishable from people and answerable to unknown parties with any range of agendas – places us into truly uncharted territory, and in a fix that will call upon a blend of government, commercial and interpersonal cooperation to see us through it.

As a start, we must establish practices of disclosure on which texts (and personalities) are AI-generated versus actual people, and expect major platforms to undertake some means of providing for and ascertaining that truth, the way that supermarkets and pharmacies are expected to stock wares that respect basic labeling conventions. Regulators can find plenty of precedent for this kind of requirement from their experiences handling falsity within native advertising or, more broadly, old-school provenance-detection frameworks for phenomena like phishing, or know-your-customer standards in banking.

While yesterday's boring but critical cybersecurity problems remain, there's unfortunately room for much more. Perhaps not just the gravity but the suddenness of these new threats will inspire consensus action over mere essays.

## JONATHAN ZITTRAIN

is the George Bemis Professor of International Law at Harvard Law School, Professor of Public Policy at the Harvard Kennedy School of Government, Professor of Computer Science at the Harvard School of Engineering and Applied Sciences, and Co-Founder of the Berkman Klein Center for Internet & Society