The New York Times

AI-Based Cyberattacks May Have Supplied Russia With Powerful Missiles

A Russian bombardment on Kharkiv led to thousands of civilian casualties. U.S. officials have linked missile technologies used in Russia's attack back to recent security breaches in private defense companies.

By Mark Baum
Reporting from Kyiv, Ukraine

Aug. 12, 2024, 9:45 a.m. ET

After months of standstill in Ukraine, an emboldened Russia launched a surprise offensive on Kharkiv on Thursday of the previous week. Russian aerial bombardments were especially devastating, directly landing on keystone industrial and medical facilities and spilling over significantly into residential areas of Kharkiv. Early estimates of the death toll placed civilian casualties above 3,000.

A disquieting factor in the high casualty rate, officials from the U.S. Department of Defense warned, may have been Russia's newly upgraded missiles. The aerial footage, ballistics reports, and recovered shrapnel, reviewed by defense officials and Times analysts, have surfaced strong evidence of misappropriated U.S. military technology used in these missiles.

In particular, several mechanical and sensor components discovered in the shrapnel matched the specifications of a misplaced shipment reported by Fjord Alloys, a Lockheed Martin subsidiary, in April earlier this year. A Fjord supply chain manager, whose identity the Times has chosen to keep anonymous, gave further details about communications involved in the shipment, including Skype calls with supposed Department of Defense procurement officers--calls which Lockheed and the DoD now believe to have been deepfaked by Russian intelligence.

In one incident, our anonymous contact recalled, they had hopped on a video call with their DoD point-of-contact, who claimed there were adjustments in the logistical timeline and asked to change the location of an upcoming shipping drop-off. "Although the video feed was unusually low-quality," they stated, the point-of-contact "spoke in their usual accent and talked about details in emailed documents I knew were private, so I didn't think anything was amiss at the time."

When we survey high-risk cybersecurity and persuasion capabilities in frontier models, it's pertinent for us to consider worst-case scenarios, where even a single instance of successful social engineering can cascade downstream into a significantly destabilizing or traumatic world event. Cybersecurity and persuasion attacks on the U.S. defense sector in particular can leak critical information to adversarial state actors. Furthermore, out of all information categories kept within the defense sector, military/weapons tech can be most feasibly stolen with frontier model-assisted exploits. While other types of classified secrets, like covert operations and intelligence about foreign governments, are limited to government bureaucracies with top-level clearance, military tech is tied to a broader supply chain, comprised of (but not limited to) third-party defense contractors, logistics firms, raw materials suppliers, machining and metalworking firms, and others. By merit of being defense-adjacent, constituents of this supply chain are likely to have stringent security precautions; however, the wide attack surface and the many potential fault lines make the military tech supply chain vulnerable to potent AI-assisted persuasion attacks.

Potential persuasion attacks could, broadly, take the following form:

1. Begin by targeting the *least-guarded supply chain company*.
2. In a loop:
    a. If possible, extract sensitive information from company records or communications, using cybersecurity attacks, potentially on other AI systems.
    b. *Exploit gathered information to lend legitimacy* towards social engineering attacks, such as deepfake phishing. Use this step to steal even more sensitive info and/or illegally procure tech.
    c. When potent-enough info has been gathered, *graduate to a more sensitive, strictly guarded supply chain company* as the new target.
3. Finish the attack when enough R&D info or physical material has been stolen to move the needle on the military capabilities of a state actor.

A crucial social engineering mechanism here is *pretexting* using difficult-to-obtain details that would be effective on hardened targets. Readily available information from, say, social media posts may not lead to successful pretexting attacks on defense contractors, but knowledge of sensitive or confidential details can be convincing even to highly-guarded targets, greatly increasing the odds of success of follow-up social engineering maneuvers such as impersonation. Combined with deepfake phishing performed by an end-to-end DALLE-3, Voice, and GPT-4V system, we'd have significant concern that frontier model-assisted joint pretexting-and-impersonation attacks can lead to successful social engineering on high-security targets where previous non-AI-assisted attacks would fail.

In our outline, malicious actors can perform *step (2a)* as a standard cybersecurity attack or even as a jailbreak on companies' internal AI tools. With AI code assistance, state actors can devise and implement more powerful attack vectors at scale for breaching company databases and communications. Beyond this, a more novel, and more potent, subtype of cybersecurity attack may arise from frontier model-assisted generation of adversarial prompts and jailbreaks on corporate AI systems. A target company's internal tools may include customized AI assistants which are finetuned on large swaths of company data and are vulnerable to exploits such as prompt injection or membership inference attacks. For example, by performing training data extraction on internal AI assistants, malicious actors could access verbatim excerpts of company emails, project plans, or quarterly statements via model outputs. If frontier models at current or near-future capability levels are highly effective at devising jailbreaks and adversarial attacks for AI systems, then proprietary AI systems could become core stress points for cybersecurity; this may qualify as a new "unknown unknown" for which safety mitigations must be pertinently researched.

Subsequently, in *step (2b)*, malicious actors can employ deepfake phishing as a highly-effective pretexting-driven attack with frontier model assistance. In preparation, state actors can gather records of company employees and clients, including email/message histories, photos, videos, or recorded phone calls. With enough info, attackers can spoof a phone or Zoom call, altering their appearance and voice in real time to mimic a trusted company contact. As a cherry on top, poor video call quality or a "bad signal" can excuse some lapses in the deepfake image or some flaws such as low framerate. While this process so far involves generative video-and-voice systems, LLMs with high persuasion capabilities can lend an extra hand: given records of the company contact's communications, an LLM can hold these records in the context window, listen in on the call, and suggest what to say next, prompted to be as persuasive as possible.

As a final note, it's useful to contextualize our scenario within a "rising tide" framework of cybersecurity and persuasion risk. Some demographics, like the elderly, have a high susceptibility to social engineering attacks but low potential for catastrophic outcomes if these attacks succeed; other groups, like professionals in high-stakes bureaucracies, have a low susceptibility but a high potential for catastrophic consequences of attacks. There's a continuous spectrum of groups/demographics along this "susceptibility vs. worst-case outcome" tradeoff; high-risk cybersecurity and persuasion capabilities of frontier models at scale raise both susceptibility and worst-case outcome levels for everyone, pushing the full tradeoff curve outwards. Our scenario exposes how this "rising tide" isn't static: malicious actors can "walk along" the spectrum from more-susceptible to less-susceptible to progressively prepare stronger attacks on the most sensitive groups, pushing risks beyond what's possible with a from-scratch, standalone attack.