Likely (tragic) Pre-Cyberwar Scenarios

Erland Wittkoetter, Ph.D., Oct 11th 2022

Software vulnerabilities are mainly found by chance. Hacking is labor-intensive and requires skills that AI could enhance. It's just a matter of time until attackers create advanced "Hacker-AI" to accelerate detection and understanding of hard-/software vulnerabilities in complex/unknown technical systems.

With incredible technical complexity already around us, developers/engineers seek tools to improve their performance: getting things done quicker without knowledge/experience in low-level details. AI techniques allow technical problems (technologies) to be turned into rules for "games". AI provides capabilities without having experts use/prepare complex details as done in pre-AI development. Acquiring knowledge or technical expertise from documentation for fixing technical problems could take hours, days, or weeks. There is a huge incentive to get operational, high-quality solutions in seconds or minutes. To avoid damage, being fast is often a requirement.

Hacker tools, particularly for reverse-code-engineering, are dangerously dual-use. Tools enhanced with AI could start with minimal assumptions/features based on Deepmind's "Reward is Enough" hypothesis. They could analyze IT-based technologies without knowing the underlying CPU/OS details and provide methods to bypass OS's low-level access-control security or detect ways to gain sysadmin rights.

Automated/AI-based hacker tools are useful in cyber-defense but more valuable in cyber-offensive. Because software-operations are invisible in our IT environment, attackers have significant first-mover advantages. Even if AI detects vulnerabilities, we still must fix them and deploy updates in follow-up steps to stop attackers.

But what if a well-resourced organization creates Hacker-AI that would find, use or create vulnerabilities in devices it analyzes? They could covertly bypass access control systems and modify software via reverse code engineering. The goals are simple: steal crypto keys or user credentials, hide attack software from all detection methods, and become irremovable from visited devices. The advantage for its operators is game-changing. They would have a super-hacker and digital ghost that could permanently occupy and dominate visited devices.

It is even conceivable that Hacker-AI could defend itself from being detected. It could sabotage new detection tools with (reasonable) error messages until it determines from simulations under its control how it must manipulate it – again, software in RAM is invisible to humans.

With hypervisor solutions, it is possible to make software irremovable on devices. Late-coming Hacker-AI could be repelled by the already occupying AI, which would make the first Hacker-AI the only one – and for its operators: permanent, global IT supremacy. In that position, a cyberwar is not required because war is practically won: communication, logistics, and supply are sabotaged covertly. With weapon systems malfunctioning, a fight is useless or self-destructive. US nuclear deterrence justifies a massive first strike in this situation, but who was it? And can command and control be maintained when commercial components are involved?

An AI with these features may require additional AI breakthroughs. Deepmind's principles and code in their AI approach are remarkably simple for what they have accomplished. So, the scary thought is, what if creating Hacker-AI is not moonshot difficult? Maybe we should ask DARPA if they made progress on their cyber challenge from August 2016: bots were automatically hacking other unknown software bots while defending themselves. This raises an even scarier question: what if Hacker-AI was already envisioned, developed, and deployed? If the result was a digital ghost, how could we know? Assuming this AI is already ubiquitous, we would likely fail to detect it by scanning harddrives from other infected systems. Could our "failure of imagination" lead to a new disaster?

Ignoring or being quiet about Hacker-AI is not helpful – although that is done by the folks who know about this threat. Governmental regulation or diplomacy/ international treaties will not remove the incentives from developing/using Hacker-AI; risks of discovery is minimal while the stakes of being first can't be any higher.

We must address this problem technically. We must react automatically to Hacker-AI. We need effective, global countermeasures in seconds or minutes, not hours, days, or weeks. Cybersecurity must become proactive, preventative, separate (i.e., independent), and redundant to a level of "Security Overkill" without users noticing it. Defenders must gain the upper hand ASAP.

We intend to start the open-source developer-community No-Go-* (https://nogostar.com) to develop that security technology while helping developers get educated so that their custom/legacy solutions benefit from the same approach to security. We believe we are not helpless in dealing with technologies leading to Hacker-AI. It is not tragic because we can do something about it.

 info@NoGoStar.com