Mandiant

M-Trends

2024 Special Report

Artificial Intelligence in Red (and Purple) Team Operations

In 1955, John McCarthy, a luminary of the fields of both computer science and cognitive science, coined the term "artificial intelligence" to be "the science and engineering of making intelligent machines." This research started as a way to make a machine behave or perform like a human, but, in the modern landscape, the meaning has evolved to encompass machines that can learn like a human. While the artificial intelligence (AI) systems imagined in popular science fiction are still quite far off, AI is currently in a period of massive growth, investment, and potential. What started as logic machines with massive decision trees has expanded to highly complex algorithms based on myriad statistics and large-scale compute power. Recently, the most topical of these algorithms are those that are designed to produce something; these systems are commonly referred to as generative AI (gen AI).

> **Targeted attack lifecycle:** The typical sequence of events taken by attackers when conducting targeted operations.

Gen AI has captured recent interest and achievements following the release of multiple gen AI tools. Gen AI now creates content on scales previously unseen, and is assisting with fields that diverge from the purely technical foundations of computer science. In cybersecurity, we have seen gen AI revolutionize the field of detection engineering, where neural networks and machine learning algorithms now form the heart of a variety of detection and response toolsets. However, an area where we have yet to see substantial adoption, yet has the potential to yield significant gains, is in the field of proactive security and red team assessments.

During a red team assessment, Mandiant experts evaluate the capabilities of a customer's security programs by simulating real-world attack scenarios. Mandiant has observed that attacker usage of AI has largely been limited to the Initial Access stage of the Targeted Attack Lifecycle. Specifically, usage has been limited to social engineering and information operations. Mandiant's Red Team has leveraged gen AI in similar fashion, seeing the greatest growth in adoption when leveraging AI to gain Initial Access to client environments.

Social Engineering Pretexts

One of the most prominent examples of gen AI usage within red team assessments is assisting with the process of generating content and media. Mandiant Red Team assessments will often include a social engineering portion, during which Mandiant is tasked with convincing clients to undertake malicious actions unknowingly. This is commonly done through text- or image-based channels, such as impersonation emails or websites. Mandiant consultants have used gen AI tooling to create initial drafts of malicious emails, as well as potential landing pages under the guise of communications that are more routine. When successful, these social engineering attempts result in Mandiant gaining access to a client network, which is often the first objective of the assessment.

However, success is not the only metric that is helpful when performing social engineering campaigns during red team assessments. By offloading the setup workflow to an AI system, Mandiant is able to gain overall increases in throughput. The faster a social engineering campaign can be set up and performed, the more potential campaigns can be completed. Instead of creating a template from scratch, for which tailoring details can be quite time consuming, gen AI can be leveraged to source social engineering pretexts more quickly.

Rapid Tool Development

Much like how gen AI can be leveraged for the creation of social engineering pretexts, gen AI has also proven to be helpful in software development across many areas of programming. Mandiant has found similar advances provided when AI is used to assist in the development of custom tooling during red team engagements. Gen AI is proving to be a capable resource when assisting with well-known algorithms and data structures, can generate code from a natural language-based description, and even integrates into popular developer environments. These capabilities and integrations provide significant value when Mandiant encounters uncommon or new applications and systems—a regular occurance during red team assessments.

In cases where environments do not fit the operational norm, Mandiant looks to operationalize as much tooling as possible to assist with achieving a variety of engagement objectives. In one scenario, Mandiant consultants used gen AI to help build a set of tools that would assist with the enumeration of accessible cloud environments to provide recommendations that improve the security posture of customer environments. Without gen AI, this process would have been much lengthier, forcing consultants to spend hours scouring related documentation as opposed to operating and delivering value. Tooling built during engagements often live on well past the close of the engagement, continuing to provide value well into the future when reused. By closing the time necessary for the initial research and creation, the value gained by repeated use increases as the tooling is formalized and adopted in future engagements.

Rapid Knowledge Acquisition

During purple team engagements, Mandiant looks to become familiar with a client's environment from the perspective of both attacker and defender. Logging, data storage, and detection stacks come in a variety of packages from off-the-shelf software to custom built detection stacks made of bespoke software. This often places a consultant in an environment where they may not be fully knowledgeable of the defensive toolkit that the customer relies on for day-to-day operations. As a result, Mandiant consultants must familiarize themselves not only with the products in use, but their potential responses to the attacks being tested.

> **Red Team:** Red teams plan and execute attacks against organizations for the purposes of identifying weaknesses.

> **Purple Team:** Purple teams foster communication and collaboration between red teams and defenders to improve incident response capabilities.

Recently, Mandiant has begun to leverage gen AI in a conversation capacity to enhance understandings of platforms and subsequently hone in on the security aspects of those platforms. Conversations with gen AIs are often iterative, with broad-stroke initial requests such as asking the AI to describe the common logging methods of a specific piece of software. This provides a launchpad for the conversation to turn to more detailed topics based on subsequent questions, which reference previous answers and public documentation. While this workflow requires vetting of answers and follow-on testing, the collaborative nature of the process provides a user with a framework and initial knowledge base off which they can work. Ultimately this has led to an ability to build a more accurate understanding of the technology stack deployed within a customer's environment, a better engagement workflow, and a better product.

Future adoption of AI for Red Teams

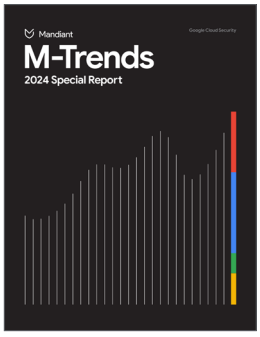
AI and large language model (LLM) development teams seek to ingrain a concept of appropriate values within a developed language model. This concept, called "AI Alignment", attempts to produce models that work to advance the designer's intended goals within the values defined, while denying requests which fall outside them. AI alignment helps provide guardrails, which limit the malicious use of an AI. While attackers have leveraged AI to become more efficient, shy of developing their own models, they have to operate within the bounds defined by the alignment or attempt to break the alignment. Google even operates their own AI-specific Red Team[1] to help find and address potential misuse of AI. However, Mandiant's red team assessments present a logical conundrum for AI alignment.

Mandiant's red team performs sanctioned malicious actions that customers have requested in order to help improve the overall security of their environments. The concept of AI Alignment places a ceiling on the level of AI adoption red teams can expect when the values encoded in the AI make it such that it will not provide answers. Conversely, red team engagements produce high-quality data, which helps drive better security outcomes for customers that can, in turn, be used to train AI models.

An exciting feature of LLMs is their ability to be fine-tuned or trained on what is known as domain specific knowledge. The majority of LLMs used by the public are generalist LLMs, meaning that the models are trained on a variety of data that covers a wide swath of different knowledge domains varying in both depth and breadth. Some LLMs are tuned for programming, whereas others might be targeted towards medical knowledge such as MedLM.[2] For the purposes of cybersecurity experts, there is Google's SecLM,[3] which is designed to provide actionable visibility into the latest threats. Tuning LLMs on specific knowledge domains requires vast amounts of specialized data within the target domain. Red teams, as professional organizations, generate and store a substantial amount of data, which could be used to train models tuned to help secure customer environments.

Resolving these logical and technical challenges will require a multi-pronged approach. Red teams will need to generate structured data on which models can be trained, and provide subject matter expertise to AI developers. Meanwhile, AI developers will have to find novel ways to pursue AI Alignment that leverages the legitimate use of malicious activity, and to properly secure access to the models that will be trained on that data. The combination of red team expertise and powerful AI leads could result in a future where red teams are considerably more effective, and organizations are better able to stay ahead of the risk posed by motivated attackers.

1 https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/
2 https://cloud.google.com/blog/topics/healthcare-life-sciences/introducing-medlm-for-the-healthcare-industry
3 https://cloud.google.com/blog/products/ai-machine-learning/gemini-for-google-cloud-is-here



Read the full report: M-Trends 2024 Special Report