Research AI and machine learning Microsoft Copilot for Security
7 min read

Microsoft AI Red Team building future of safer AI

By Ram Shankar Siva Kumar, Microsoft AI Red Team Lead

August 7, 2023

An essential part of shipping software securely is red teaming. It broadly refers to the practice of emulating real-world adversaries and their tools, tactics, and procedures to identify risks, uncover blind spots, validate assumptions, and improve the overall security posture of systems. Microsoft has a rich history of red teaming emerging technology with a goal of proactively identifying failures in the technology. As AI systems became more prevalent, in 2018, Microsoft established the AI Red Team: a group of interdisciplinary experts dedicated to thinking like attackers and probing AI systems for failures.

We're sharing best practices from our team so others can benefit from Microsoft's learnings. These best practices can help security teams proactively hunt for failures in AI systems, define a defense-in-depth approach, and create a plan to evolve and grow your security posture as generative AI systems evolve.

The practice of AI red teaming has evolved to take on a more expanded meaning: it not only covers probing for security vulnerabilities, but also includes probing for other system failures, such as the generation of potentially harmful content. AI systems come with new risks, and red teaming is core to understanding those novel risks, such as prompt injection and producing ungrounded content. AI red teaming is not just a nice to have at Microsoft; it is a cornerstone to responsible AI by design: as Microsoft President and Vice Chair, Brad Smith, announced, Microsoft recently committed that all high-risk AI systems will go through independent red teaming before deployment.

The goal of this blog is to contextualize for security professionals how AI red teaming intersects with traditional red teaming, and where it differs. This, we hope, will empower more organizations to red team their own AI systems as well as provide insights into leveraging their existing traditional red teams and AI teams better.

Red teaming helps make AI implementation safer

Over the last several years, Microsoft's AI Red Team has continuously created and shared content to empower security professionals to think comprehensively and proactively about how to implement AI securely. In October 2020, Microsoft collaborated with MITRE as well as industry and academic partners to develop and release the Adversarial Machine Learning Threat Matrix, a framework for empowering security analysts to detect, respond, and remediate threats. Also in 2020, we created and open sourced Microsoft Counterfit, an automation tool for security testing AI systems to help the whole industry improve the security of AI solutions. Following that, we released the AI security risk assessment framework in 2021 to help organizations mature their security practices around the security of AI systems, in addition to updating Counterfit. Earlier this year, we announced additional collaborations with key partners to help organizations understand the risks associated with AI systems so that organizations can use them safely, including the integration of Counterfit into MITRE tooling, and collaborations with Hugging Face on an AI-specific security scanner that is available on GitHub.

Security-related AI red teaming is part of a larger responsible AI (RAI) red teaming effort that focuses on Microsoft's AI principles of fairness, reliability and safety, privacy and security, inclusiveness, transparency, and accountability. The collective work has had a direct impact on the way we ship AI products to our customers. For instance, before the new Bing chat experience was released, a team of dozens of security and responsible AI experts across the company spent hundreds of hours probing for novel security and responsible AI risks. This was in *addition* to the regular, intensive software security practices followed by the team, as well as red teaming the base GPT-4 model by RAI experts in advance of developing Bing Chat. Our red teaming findings informed the systematic measurement of these risks and built scoped mitigations before the product shipped.

Guidance and resources for red teaming

AI red teaming generally takes place at two levels: at the base model level (e.g., GPT-4) or at the application level (e.g., Security Copilot, which uses GPT-4 in the back end). Both levels bring their own advantages: for instance, red teaming the model helps to identify early in the process how models can be misused, to scope capabilities of the model, and to understand the model's limitations. These insights can be fed into the model development process to improve future model versions but also get a jump-start on which applications it is most suited for. Application-level AI red teaming takes a system view, of which the base model is one part. For instance, when AI red teaming Bing Chat, the entire search experience powered by GPT-4 was in scope and was probed for failures. This helps to identify failures beyond just the model-level safety mechanisms, by including the overall application specific safety triggers.

Together, probing for both security and responsible AI risks provides a single snapshot of how threats and even benign usage of the system can compromise the integrity, confidentiality, availability, and accountability of AI systems. This combined view of security and responsible AI provides valuable insights not just in proactively identifying issues, but also to understand their prevalence in the system through measurement and inform strategies for mitigation. Below are key learnings that have helped shape Microsoft's AI Red Team program.

1. **AI red teaming is more expansive**. AI red teaming is now an umbrella term for probing both security and RAI outcomes. AI red teaming intersects with traditional red teaming goals in that the security component focuses on model as a vector. So, some of the goals may include, for instance, to steal the underlying model. But AI systems also inherit new security vulnerabilities, such as prompt injection and poisoning, which need special attention. In addition to the security goals, AI red teaming also includes probing for outcomes such as fairness issues (e.g., stereotyping) and harmful content (e.g., glorification of violence). AI red teaming helps identify these issues early so we can prioritize our defense investments appropriately.
2. **AI red teaming focuses on failures from both malicious and benign personas**. Take the case of red teaming new Bing. In the new Bing, AI red teaming not only focused on how a malicious adversary can subvert the AI system via security-focused techniques and exploits, but also on how the system can generate problematic and harmful content when regular users interact with the system. So, unlike traditional security red teaming, which mostly focuses on only malicious adversaries, AI red teaming considers broader set of personas and failures.
3. **AI systems are constantly evolving.** AI applications routinely change. For instance, in the case of a large language model application, developers may change the metaprompt (underlying instructions to the ML model) based on feedback. While traditional software systems also change, in our experience, AI systems change at a faster rate. Thus, it is important to pursue multiple rounds of red teaming of AI systems and to establish systematic, automated measurement and monitor systems over time.
4. **Red teaming generative AI systems requires multiple attempts**. In a traditional red teaming engagement, using a tool or technique at two different time points on the same input, would always produce the same output. In other words, generally, traditional red teaming is deterministic. Generative AI systems, on the other hand, are probabilistic. This means that running the same input twice may provide different outputs. This is by design because the probabilistic nature of generative AI allows for a wider range in creative output. This also makes it tricky to red teaming since a prompt may not lead to failure in the first attempt, but be successful (in surfacing security threats or RAI harms) in the succeeding attempt. One way we have accounted for this is, as Brad Smith mentioned in his blog, to pursue multiple rounds of red teaming in the same operation. Microsoft has also invested in automation that helps to scale our operations and a systemic measurement strategy that quantifies the extent of the risk.
5. **Mitigating AI failures requires defense in depth.** Just like in traditional security where a problem like phishing requires a variety of technical mitigations such as hardening the host to smartly identifying malicious URIs, fixing failures found via AI red teaming requires a defense-in-depth approach, too. This involves the use of classifiers to flag potentially harmful

content to using metaprompt to guide behavior to limiting conversational drift in conversational scenarios.

Building technology responsibly and securely is in Microsoft's DNA. Last year, Microsoft celebrated the 20-year anniversary of the Trustworthy Computing memo that asked Microsoft to deliver products "as available, reliable and secure as standard services such as electricity, water services, and telephony." AI is shaping up to be the most transformational technology of the 21st century. And like any new technology, AI is subject to novel threats. Earning customer trust by safeguarding our products remains a guiding principle as we enter this new era – and the AI Red Team is front and center of this effort. We hope this blog post inspires others to responsibly and safely integrate AI via red teaming.

Resources

AI red teaming is part of the broader Microsoft strategy to deliver AI systems securely and responsibly. Here are some other resources to provide insights into this process:

- For customers who are building applications using Azure OpenAI models, we released a guide to help them assemble an AI red team, define scope and goals, and execute on the deliverables.
- For security incident responders, we released a bug bar to systematically triage attacks on ML systems.
- For ML engineers, we released a checklist to complete AI risk assessment.
- For developers, we released threat modeling guidance specifically for ML systems.
- For anyone interested in learning more about responsible AI, we've released a version of our Responsible AI Standard and Impact Assessment, among other resources.
- For engineers and policymakers, Microsoft, in collaboration with Berkman Klein Center at Harvard University, released a taxonomy documenting various machine learning failure modes.
- For the broader security community, Microsoft hosted the annual Machine Learning Evasion Competition.
- For Azure Machine Learning customers, we provided guidance on enterprise security and governance.

Contributions from Steph Ballard, Forough Poursabzi, Amanda Minnich, Gary Lopez Munoz, and Chang Kawaguchi.

Explore Microsoft products

Windows 11 apps

Microsoft Store

Account profile

Download Center

Microsoft Store support

Returns

Order tracking

Certified Refurbished

Microsoft Store Promise

Flexible Payments

Education

Microsoft in education

Devices for education

Microsoft Teams for Education

Microsoft 365 Education

How to buy for your school

Educator training and development

Deals for students and parents

Azure for students

Business

Microsoft Cloud

Microsoft Security

Dynamics 365

Microsoft 365

Microsoft Power Platform

Microsoft Teams

Copilot for Microsoft 365

Small Business

Developer & IT

Azure

Developer Center

Documentation

Microsoft Learn

Microsoft Tech Community

Azure Marketplace

AppSource

Visual Studio

Company

Careers

About Microsoft

Company news

Privacy at Microsoft

Investors

Diversity and inclusion

Accessibility

Sustainability

English (United States)

Your Privacy Choices

Consumer Health Privacy