Dear esteemed members of NIST,

Thank you for your valuable work producing the AI Risk Management Framework, as well as for this opportunity to provide feedback; particularly on effective AI red-teams.

My recommendations are based on a decade of experience in privacy measurement and assurance in government, academia, and tech companies. I have also built and run red-team exercises for privacy, and have seen how red-team testing for human rights and safety has particular requirements. Therefore, I provide three high-level recommendations about AI red-teams:

1. Recommendations on a clear definition of "AI red-team" based on what works for privacy red-teams and therefore may be needed for explainability, bias, or other AI risks.
2. Current best practices through a four-step process for running an AI red-team.
3. A discussion on prioritizing vulnerabilities found by an AI red-ream.

AI red-team may differ from security red-teams in three key ways. Alternatively, AI testing should include these elements.

Without a clear definition of what an AI "red-team" is, it is difficult to discuss the pros and cons and how to run an exercise effectively. It is also difficult to describe what other testing should complement AI red-teaming.

I assume an "AI red-team" is a group of human experts who run non-automated tests on an AI system. I would like to highlight some specific elements of testing that are needed for AI red-teaming. AI red-teams should include clear-box, non-adversarial testing. Furthermore, AI red-teams may be useful even when no "blue-team" exists. These may be different from the common understanding of security red-teams. My reasoning is based on experience with privacy.

1. AI privacy tests should include clear-box testing. Clear-box testing means the AI red-team has access to or clarity on the training data, the model itself, or other information about how the model was built. Closed-box testing implies the AI red-teaming only has access to the outputs of the system (the "box" of the system is closed to testers). Effective measures for protecting privacy in AI could only be tested in a clear-box method. For example, protections such as applying differential privacy to the training data are immeasurable in the output alone. Therefore, outcome-only AI testing will be less effective for privacy, and I suspect for other elements of AI risk.
2. AI testing need not assume malicious activities. An AI red-team can emulate a motivated, external adversary, but testing can and should also include other types of flaws. Vulnerabilities that are possible without any malicious actor often include "own-goals" introduced by the organization responsible for the model due to mistakes, poor planning, or unawareness of the risk. These vulnerabilities should be included in AI testing efforts.

3. AI red-team testing need not include a blue team. Red-team testing in security and military exercises is often designed to test the "blue-team", or to test the defense and detection capability. For many of the AI risks around human rights, it is unclear who or what the "blue-team" would be and how they would be measured. Therefore, AI testing must include a scan for all vulnerabilities, and should not be limited to testing defense and detection.

AI red teams should follow a four-step process: model, plan, run the exercise, and communicate.

The goal of these four steps is to highlight how AI red-teams should protect people, protect the red-team testers, and provide effective practical outcomes that can be addressed and fixed.

1. Threat model. Define what risks or vulnerabilities will be included in the red-team exercise. Understand the potential concerns, the context, and the harms that will be investigated.
2. Plan for the exercise. There should be clear rules of engagement. Prepare tools to run the exercise. Any red-team exercises with impact or changes outside of the red-team members should be part of an ethical review (perhaps similar to an IRB for human-subjects tests) before starting. Also consider whether the red-team testers themselves will be subject to traumatic or violent results, and plan how to mitigate these harms.
3. Run the exercise following a pre-defined scope with clear objectives. Each step of the exercise should be logged in detail so that it can be replicated or compared afterward. This logging may be similar to a chemistry lab book; even attempts that fail should be recorded. I've seen the best results when red-team members can communicate with each other during the exercise to share ideas and overcome roadblocks.
4. Communicate and remediate the vulnerabilities found. Red-team testers need to be creative when it comes to understanding the AI risks and how to uncover them. Red-teams should communicate in writing or in presentations to decision-makers what the risks were, why they mattered, and potentially what a fix would look like. This may mean that an AI red-team also needs to have a specialist in remediation communication and organization.

AI red-teams should have access to authority, should not have a conflict of interest with AI safety due to work responsibilities, and should not be terminated for finding vulnerabilities.

Fixing multiple vulnerabilities will be hard because there are no clear standards to prioritize the different types of vulnerabilities an AI red-team might find.

In this section, I assume that vulnerabilities can outnumber the resources or ability to fix them simultaneously. Deciding which vulnerability to address first can require time-consuming discussions that delay or block the remediation work.

There is no existing industry-wide standard to prioritize vulnerabilities caused by different types of AI risks. Even within one category of risk, such as privacy, prioritizing fixes is hard. The privacy community does not have a vulnerability scoring standard that is generally agreed upon.

The taxonomy of AI risks includes several that cannot all be simultaneously optimized. For example, accuracy may be at odds with privacy. Privacy protections in datasets include adding noise, allowing plausible deniability, or removing some outlying pieces of data that could identify individuals. All these privacy-protective mechanisms can reduce accuracy. Therefore, if an AI organization is faced with a choice between accuracy or privacy, it is unclear which provides the most safety to society.

A scoring system could help organizations prioritize fixes. NIST may be the right place to develop an AI vulnerability score.

In the meantime, a practical first step for organizations is to run focused red-team exercises with clear goals and priorities. Managing and mapping vulnerabilities at an early stage will be more effective than running an AI red team on all possible risks and then trying to prioritize the vulnerabilities. AI red-team exercises should be based on a transparent threat model process, a clear statement of success, and resources for remediation.

Thank you for your consideration. I hope these are useful in the practical development of efficient and effective AI red-teaming.

Thank you,
Rebecca Balebako

www.privacyengineer.ch