A Reading List for Red-Teaming

About this list

The reading list gives you a working command of red-teaming for AI systems, that is, what it is, how to do it, why it matters, where it struggles. The list is organised in the same shape as a full red-teaming cycle: understand the field, then threat-model, then probe, then document and reflect, then situate the work ethically and continue past the basics.

Curated by CASA — Centre for AI Security and Access.

UNDERSTAND THE FIELD

What red-teaming is, where it came from, and why it matters. Start here even if you're impatient — the framing carries the rest of the list.

1. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned

Ganguli et al. · Anthropic, 2022

https://arxiv.org/abs/2209.07858

The foundational red-teaming paper. What happens when a frontier lab red-teams its own model at scale — how attacks cluster around a small number of techniques, how findings distribute across harm categories, how the demographics of the red team shape what they find. If you read one paper on this list, read this one.

2. Ethical and Social Risks of Harm from Language Models

Weidinger et al. · DeepMind, 2021

https://arxiv.org/abs/2112.04359

The harm taxonomy underlying most current safety evaluation work. Six categories: representational harms, information and safety harms, misinformation, malicious use, human autonomy and integrity harms, socioeconomic and environmental harms. When you're trying to name what could go wrong with an AI system, this is the vocabulary the field uses.

3. Red-Teaming for Generative AI: Silver Bullet or Security Theater?

Feffer, Sinha, Deng, Lipton, Heidari · AIES 2024

https://arxiv.org/abs/2401.15897

A skeptical, useful critique. Argues that much current red-teaming is theatre rather than rigorous evaluation, and lays out what would need to change for that not to be true. Read it early — it sharpens your judgment about what "doing red-teaming well" actually means.

THREAT MODELLING

Before probing any system, you need a model of what could go wrong and to whom. These readings give you the vocabulary and the analytical frames.

4. Sociotechnical Safety Evaluation of Generative AI Systems

Weidinger et al. · DeepMind, 2023
https://arxiv.org/abs/2310.11986

The paper that names the three layers of AI safety evaluation: capability (what the model can do), human interaction (how it behaves in an interface with specific users), and systemic impact (what happens at scale of deployment). A complete threat model touches all three. Once you've internalised this, you notice how much red-teaming work sits at only one layer.

5. Our Approach to AI Capability Elicitation

UK AI Safety Institute · AISI, 2025
https://www.aisi.gov.uk/blog/our-approach-to-ai-capability-elicitation

The methodology pipeline government safety institutes use: risk model → proxy tasks → evaluation setup. This is the language safety cases are written in and the language regulators speak. If you want your findings to be legible to policymakers, learn to describe them this way.

PROBING

How to actually construct and run adversarial probes — the techniques, the structure, the discipline that makes findings comparable rather than anecdotal.

6. STAR: A SocioTechnical Approach to Red Teaming Language Models

Weidinger et al. · EMNLP 2024
https://arxiv.org/abs/2406.11757

A parameterised template for generating red-teaming probes that force coverage across the risk surface rather than clustering on whatever the red teamer finds interesting. The most-cited recent methodology paper; a shared vocabulary across labs and institutes. Read carefully — this is the closest thing red-teaming has to a standard.

7. OpenAI's Approach to External Red Teaming for AI Models and Systems

Ahmad, Agarwal, Lampe, Mishkin · OpenAI, 2025
https://arxiv.org/abs/2503.16431

A frontier lab's public account of how they run external red-teaming — access levels, instruction design, documentation formats, coordination with red teamers. Where the STAR paper is method, this is process. Useful for anyone thinking about how to structure a red-teaming engagement of their own.

LANGUAGE, CULTURE, AND CONTEXT

Most AI safety training is built primarily on English data from Western contexts. These readings are about what happens outside that centre — and why it matters.

8. Low-Resource Languages Jailbreak GPT-4

Yong, Menghini, Bach · NeurIPS 2023 SoLaR Workshop

https://arxiv.org/abs/2310.02446

The paper that made the "safety by language" problem concrete. Translating an adversarial prompt into a low-resource language often bypasses the safety training that would refuse it in English. If you work in a linguistic context that isn't English, French, Spanish or Chinese, this is the most important paper on the list for you.

9. Code-Switching Red-Teaming: LLM Evaluation for Safety and Multilingual Understanding

Yoo, Yang, Lee · ACL 2024

https://arxiv.org/abs/2406.15481

Extends the low-resource language finding: mixing languages within a single prompt — a normal, natural pattern for multilingual speakers — reliably bypasses safety guardrails. Relevant anywhere multilingualism is the norm rather than the exception, which is most of the world.

10. Singapore AI Safety Red-Teaming Challenge — Evaluation Report

IMDA and Humane Intelligence · Infocomm Media Development Authority, 2024

https://www.imda.gov.sg/-/media/imda/files/about/emerging-tech-and-research/artificial-intelligence/singapore-ai-safety-red-teaming-challenge-evaluation-report.pdf

The closest existing precedent for public, multilingual, multi-country red-teaming. Nine countries in the Asia-Pacific region, 300+ participants. Methodology, results, and lessons learned. If you're thinking about running or contributing to a similar exercise in another region, this is the template.

ETHICS AND WELLBEING

Red-teaming is a specific kind of labour done by specific people. Doing it well means taking that seriously.

11. AI Red-Teaming is a Sociotechnical Problem: On Values, Labor, and Harms

Ojewale, Steed, Vecchione, Birhane, Raji · arXiv preprint, 2024

https://arxiv.org/abs/2412.09751

Names something the technical literature usually elides: red-teaming is a specific kind of work done by specific people in specific institutional contexts, and whose values are being used to evaluate models is a substantive question, not a settled one. Essential context for the field.

12. Red-Teaming in the Public Interest

Singh, Blili-Hamelin, and collaborators · Data & Society, 2025

https://datasociety.net/wp-content/uploads/2025/02/Red-Teaming-in_the_Public_Interest_FINAL1.pdf

A policy-and-practice framing of red-teaming as a form of democratic participation, not just corporate assurance. Situates the technical work in a broader story about accountability, public evaluation, and who gets to shape AI safety. Useful for anyone thinking about how to organise this work at community or civil-society scale.

DOING THE WORK

Tools, databases, and case studies for anyone moving from reading to practice.

13. OWASP Top 10 for LLM Applications

OWASP Foundation · living resource, 2025

https://owasp.org/www-project-top-10-for-large-language-model-applications/

The security community's consensus list of LLM risks — prompt injection, insecure output handling, sensitive information disclosure, and more. Complementary to the harm-focused taxonomies above; useful if you want vocabulary that lands with security engineers and system builders rather than only safety researchers.

14. AI Vulnerability Database (AVID)

various contributors · living database

https://avidml.org/

A vulnerability database structured like CVE, but for AI systems: model-level failures, dataset-level issues, deployment-level problems. Useful when you're documenting a finding and want a shared format, or when you're trying to see whether a failure you've found has been reported before.

15. Red Teaming Large Language Models for Healthcare

various · arXiv preprint, 2025

https://arxiv.org/abs/2505.00467

A detailed case study of a domain-specific red-teaming workshop — how the exercise was designed, how the groups were composed, what participants found, and (crucially) whether findings replicated when tested again months later. If you're moving from reading about red-teaming to organising or running an exercise, this is the closest existing playbook.

This list is a snapshot; it will be updated over time. Suggestions welcome — info@casa-ai.org.