COURSE SYLLABUS

AI Red Teaming and Assurance

Twelve hours live on AI red teaming and assurance: running an engagement, matching technique to attack surface, building an auditable assurance case, and testing agentic systems. No lab needed.

Group classes (Live Online) and onsite training are available for this course. For more information, email corporate@nobledesktop.com or visit: https://www.nobledesktop.com/classes/ai-red-teaming-and-assurance



hello@nobledesktop.com • (212) 226-4149

Course Outline

Module 1

- Foundations: What Is AI Red Teaming, and Where Does It Fit in AI Assurance?
- What red teaming is and is not
- Where it sits in the AI assurance lifecycle
- The range of assurance activities
- The guidance the course draws on
- Who is involved

Module 2

- Planning and Scoping a Red-Team Engagement
- The engagement lifecycle
- Scoping to risk classification
- Setting red-team objectives
- Rules of engagement
- Team composition and deconfliction

Module 3

- Adversarial Testing Techniques for Machine Learning and Generative AI
- Matching technique to attack surface
- Manual versus automated and tool-assisted red teaming
- Generative AI probing techniques
- A preview of testing agentic and tool-using systems
- Telling red-team findings apart from ordinary bugs

Module 4

- AI Assurance Frameworks: TEVV, Assurance Cases, and Documentation
- Test, Evaluation, Verification and Validation

- The assurance case as claim, argument and evidence
- Model cards and system cards
- Independent evaluation and measurement science
- From assurance documentation to a risk-acceptance decision

Module 5

- From Findings to Fixes: Reporting, Remediation, and Human Oversight
- Structuring a findings report
- Prioritizing by impact rather than technical complexity
- Remediation and retesting
- Human oversight as a mitigation category
- When a finding becomes an incident

Module 6

- Red Teaming and Assurance for Generative and Agentic AI; Capstone Scenario
- Foundation models and misuse-risk testing
- Agentic AI and the expansion of the attack surface from output to action
- Why red teaming and assurance are recurring rather than one-time
- A capstone scenario