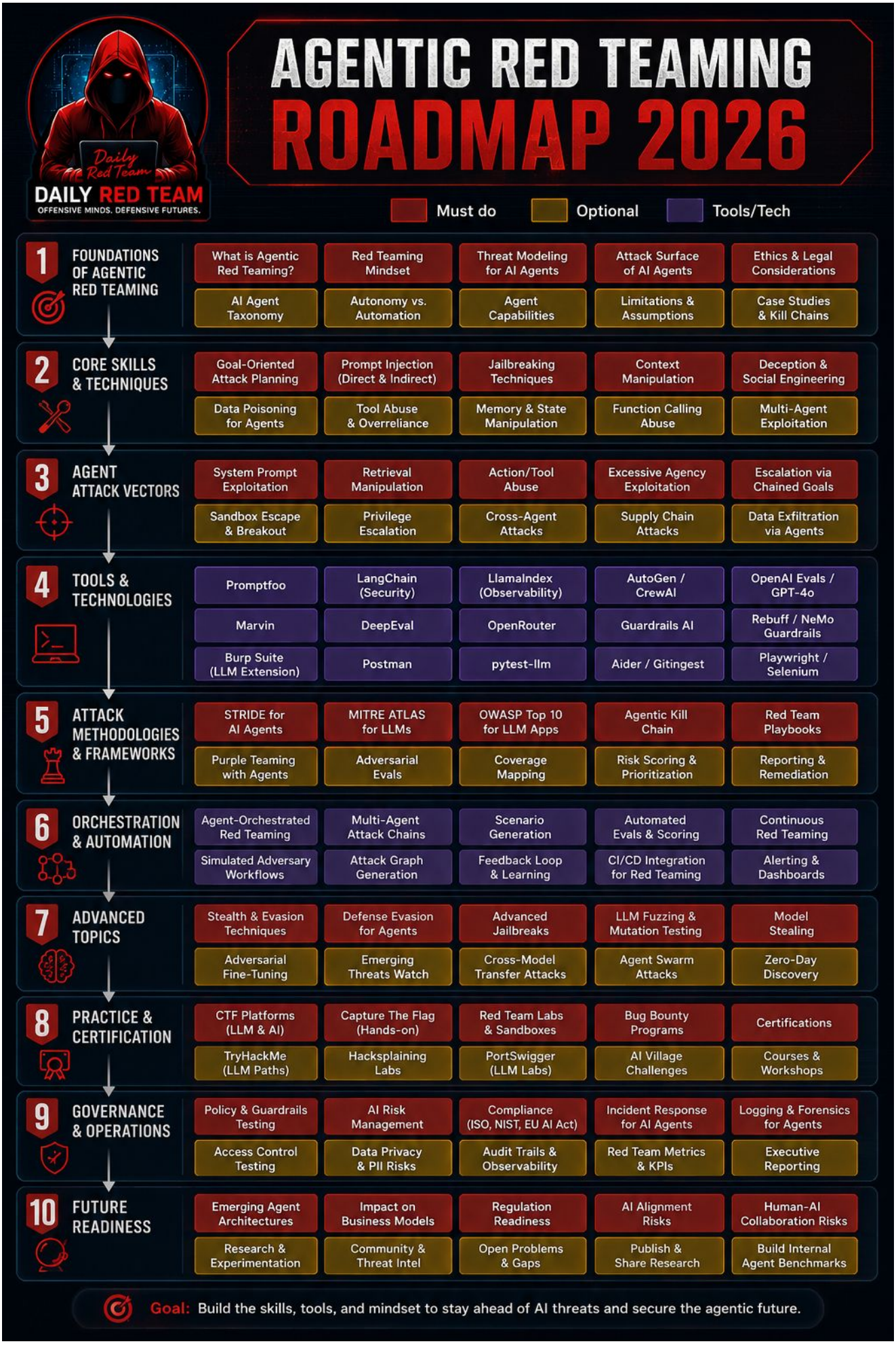
AI


Daily Red Team
DAILY RED TEAM
OFFENSIVE MINDS. DEFENSIVE FUTURES.
AGENTIC RED TEAMING
ROADMAP 2026
Must do
Optional
Tools/Tech
1 FOUNDATIONS OF AGENTIC RED TEAMING
What is Agentic Red Teaming?
Red Teaming Mindset
Threat Modeling for AI Agents
Attack Surface of AI Agents
Ethics & Legal Considerations
AI Agent Taxonomy
Autonomy vs. Automation
Agent Capabilities
Limitations & Assumptions
Case Studies & Kill Chains
2 CORE SKILLS & TECHNIQUES
Goal-Oriented Attack Planning
Prompt Injection (Direct & Indirect)
Jailbreaking Techniques
Context Manipulation
Deception & Social Engineering
Data Poisoning for Agents
Tool Abuse & Overreliance
Memory & State Manipulation
Function Calling Abuse
Multi-Agent Exploitation
3 AGENT ATTACK VECTORS
System Prompt Exploitation
Retrieval Manipulation
Action/Tool Abuse
Excessive Agency Exploitation
Escalation via Chained Goals
Sandbox Escape & Breakout
Privilege Escalation
Cross-Agent Attacks
Supply Chain Attacks
Data Exfiltration via Agents
4 TOOLS & TECHNOLOGIES
Promptfoo
LangChain (Security)
LlamaIndex (Observability)
AutoGen / CrewAI
OpenAI Evals / GPT-4o
Marvin
DeepEval
OpenRouter
Guardrails AI
Rebuff / NeMo Guardrails
Burp Suite (LLM Extension)
Postman
pytest-llm
Aider / Gitingest
Playwright / Selenium
5 ATTACK METHODOLOGIES & FRAMEWORKS
STRIDE for AI Agents
MITRE ATLAS for LLMs
OWASP Top 10 for LLM Apps
Agentic Kill Chain
Red Team Playbooks
Purple Teaming with Agents
Adversarial Evals
Coverage Mapping
Risk Scoring & Prioritization
Reporting & Remediation
6 ORCHESTRATION & AUTOMATION
Agent-Orchestrated Red Teaming
Multi-Agent Attack Chains
Scenario Generation
Automated Evals & Scoring
Continuous Red Teaming
Simulated Adversary Workflows
Attack Graph Generation
Feedback Loop & Learning
CI/CD Integration for Red Teaming
Alerting & Dashboards
7 ADVANCED TOPICS
Stealth & Evasion Techniques
Defense Evasion for Agents
Advanced Jailbreaks
LLM Fuzzing & Mutation Testing
Model Stealing
Adversarial Fine-Tuning
Emerging Threats Watch
Cross-Model Transfer Attacks
Agent Swarm Attacks
Zero-Day Discovery
8 PRACTICE & CERTIFICATION
CTF Platforms (LLM & AI)
Capture The Flag (Hands-on)
Red Team Labs & Sandboxes
Bug Bounty Programs
Certifications
TryHackMe (LLM Paths)
Hacksplaining Labs
PortSwigger (LLM Labs)
AI Village Challenges
Courses & Workshops
9 GOVERNANCE & OPERATIONS
Policy & Guardrails Testing
AI Risk Management
Compliance (ISO, NIST, EU AI Act)
Incident Response for AI Agents
Logging & Forensics for Agents
Access Control Testing
Data Privacy & PII Risks
Audit Trails & Observability
Red Team Metrics & KPIs
Executive Reporting
10 FUTURE READINESS
Emerging Agent Architectures
Impact on Business Models
Regulation Readiness
AI Alignment Risks
Human-AI Collaboration Risks
Research & Experimentation
Community & Threat Intel
Open Problems & Gaps
Publish & Share Research
Build Internal Agent Benchmarks
Goal: Build the skills, tools, and mindset to stay ahead of AI threats and secure the agentic future.

From:
https://wiki.fortier-family.com/ - **Warnaud's Wiki**

Permanent link:
https://wiki.fortier-family.com/cybersecurity/ai?rev=1789977015

Last update: **2026/09/21 09:50**