SWISS CYBER AI RESEARCH · WHITEPAPER

Cybersecurity and Artificial Intelligence

A synthesis of the leading research and standards on securing AI systems, the dual-use of AI in cyber offence and defence, and the regulatory landscape — with practical guidance for CISOs, Data Protection Officers and GDPR compliance.

Author: G. Moresi
SwissCyberAI.org — Baar, Canton Zug, Switzerland

DOCUMENT TYPE
Research paper / literature synthesis

VERSION
1.0 · 2026 (EN)

AUDIENCE
CISOs · DPOs · Security & Governance

CLASSIFICATION
Public / Internal

SwissCyberAI.org
swisscyberai.org · cyberriskevaluator.com

2026
E N

Abstract

ABSTRACT

Artificial intelligence has become simultaneously a critical asset, a new attack surface, and a powerful instrument for both attackers and defenders. This paper synthesises the current state of knowledge at the intersection of cybersecurity and AI, drawing on the principal standards and peer-reviewed literature — including the NIST AI Risk Management Framework and its adversarial machine-learning taxonomy, the OWASP Top 10 for LLM and Agentic Applications, MITRE ATLAS, the ENISA Threat Landscape, the EU AI Act, and the foundational academic work on adversarial examples, data poisoning and privacy-inference attacks.

We organise the field into three lenses: **securing AI** (attacks against models, data and pipelines), **AI for offence** (the weaponisation of generative and agentic AI), and **AI for defence** (detection, triage and augmentation). We then map these technical risks onto the regulatory and data-protection regime — the EU AI Act's Article 15 security obligations and the GDPR as interpreted by the EDPB's Opinion 28/2024 — and translate the whole into concrete, role-specific guidance for CISOs and Data Protection Officers.

Keywords: AI security, adversarial machine learning, LLM security, prompt injection, data poisoning, privacy attacks, GDPR, EU AI Act, DPO, CISO, governance.

Contents

If page numbers appear blank, select all (Ctrl+A) and press F9 to update the table of contents.

1 Introduction

The convergence of cybersecurity and artificial intelligence has moved from a research curiosity to an operational reality. AI systems are now embedded across financial services, healthcare, public administration and critical infrastructure, and with that adoption comes a threefold security problem. AI is a **new asset to protect**, because models, training data and pipelines can be attacked in ways that classical controls do not anticipate. AI is a **new weapon**, because generative and agentic systems lower the cost and skill threshold of offensive operations. And AI is a **new defensive tool**, because the same capabilities accelerate detection, triage and response.

This duality is visible in the field data. The European Union Agency for Cybersecurity analysed 4,875 incidents between July 2024 and June 2025 and found phishing to be the dominant intrusion vector, with more than four-fifths of observed phishing and social-engineering activity now AI-assisted.[11] On the research side, an arXiv search returned over 11,000 adversarial-machine-learning papers since 2021,[3] and demonstrations in 2024–2025 established that language-model agents can autonomously exploit certain known vulnerabilities and drive penetration-testing workflows.[21][22][23]

Scope and method. This paper is a structured synthesis of the most influential standards and peer-reviewed sources rather than a single new experiment. It integrates the governance frameworks (NIST, ISO, OWASP, MITRE, ENISA), the regulatory instruments (the EU AI Act and the GDPR), and the academic literature on adversarial and privacy attacks, and distils them into role-specific guidance. Throughout the paper, coloured notes flag concrete takeaways for the two roles most accountable for AI risk in practice — the **CISO** and the **Data Protection Officer (DPO)** — together with specific **GDPR** considerations.

2 The standards and framework landscape

A practitioner entering this field confronts a crowded map of frameworks. They are complementary rather than competing: governance frameworks tell an organisation how to organise, threat taxonomies tell it what to defend against, and regulation tells it what is mandatory. Table 1 orients the reader before the technical detail that follows.

Table 1 — The principal reference points for cybersecurity and AI.

Instrument	Type	Contribution
NIST AI RMF 1.0 (AI 100-1)	Governance	Voluntary, lifecycle framework organised around four functions — Govern, Map, Measure, Manage.[1]
NIST Generative AI Profile (AI 600-1)	Governance	GenAI-specific risks and a catalogue of over 400 mitigation actions mapped to the RMF functions.[2]
NIST AML Taxonomy (AI 100-2e2025)	Threat taxonomy	Formal taxonomy of evasion, poisoning and privacy attacks on predictive AI, plus prompt injection, indirect injection, RAG poisoning and backdoors on generative AI.[3]
ISO/IEC 42001:2023	Management system	Certifiable management system for AI (an “AIMS”), analogous to ISO 27001 for information security.[4]
ISO/IEC 27001:2022 · 23894 · 27090/27091	Management / guidance	ISMS baseline; AI risk-management guidance; emerging guidance on AI security and AI privacy.[5][6][7]
OWASP Top 10 — LLM & Agentic	Threat taxonomy	The most widely referenced enumeration of LLM and agentic application risks.[8][9]
MITRE ATLAS	Threat knowledge base	Adversary tactics and techniques for AI systems, modelled on ATT&CK.[10]
ENISA Threat Landscape	Threat intelligence	Annual, incident-based view of the EU threat environment, including AI-enabled threats.[11][12]
EU AI Act — Reg. 2024/1689	Regulation	Risk-tiered obligations; Article 15 mandates accuracy, robustness and cybersecurity for high-risk AI.[13]
GDPR + EDPB Opinion 28/2024	Regulation / guidance	Data-protection law and its authoritative interpretation for AI model training and deployment.[14][15]

ORIENTATION

A useful mental model: **ISO 42001** and the **NIST AI RMF** give you the governance skeleton; **OWASP**, **MITRE ATLAS** and the **NIST AML taxonomy** give you the threat catalogue that populates your risk assessments; and the **EU AI Act** and **GDPR** convert selected items into legal obligations.

3 Securing AI: a taxonomy of attacks against AI systems

Attacks against AI systems can be organised by the attacker's goal — to cause a misclassification, to corrupt the model, to extract private information, or to subvert an AI application's behaviour. The NIST adversarial-machine-learning taxonomy provides the canonical structure for predictive systems, and extends it with generative-AI classes; the OWASP Top 10 provides the application-layer view.[3][8] Table 2 summarises the landscape.

3.1 Evasion and adversarial examples

Evasion attacks perturb an input at inference time so that a model produces an attacker-chosen output while the input appears unchanged to a human. The foundational result — that imperceptible perturbations can reliably fool deep neural networks — established adversarial examples as a durable, transferable weakness of learned systems.[20] In security-relevant settings this manifests as bypassing malware, spam or fraud classifiers, or defeating biometric and content-moderation models. Defences include adversarial training, input transformation and robustness testing, but no defence is complete, and the arms race is ongoing.[3]

3.2 Poisoning, backdoors and the AI supply chain

Where evasion targets inference, poisoning targets training. By corrupting a fraction of the training data — in some settings a vanishingly small fraction — an attacker can degrade accuracy or, more insidiously, install a backdoor that behaves normally until a trigger is present.[3] This risk now extends across the AI supply chain: ENISA documents poisoned machine-learning models hosted on public platforms, trojanised packages, and a "Rules File Backdoor" that injects malicious instructions into the configuration files of AI coding assistants, as well as "slopsquatting," in which attackers register package names that language models are prone to hallucinate.[11]

> **CISO HINT**
>
> Treat models, datasets and prompts as first-class supply-chain artefacts. Maintain an **AI bill of materials (AI-BOM)**, verify the provenance and integrity of third-party models and datasets, pin and scan dependencies used by coding assistants, and extend software-composition analysis to the ML pipeline. Poisoning is cheap for the attacker and hard to detect after the fact — provenance controls are your highest-leverage investment.

3.3 Privacy attacks and model theft

A trained model can leak information about the data it was trained on. Three attack families are well established in the literature. **Membership inference** determines whether a specific record was in the training set, exploiting the tendency of models to be more confident on data they have seen; the seminal black-box construction used "shadow models" and remains the reference attack.[17] **Model inversion** reconstructs sensitive input features from model outputs — famously recovering recognisable face images from a facial-recognition model given only a name and confidence scores.[18] **Training-data extraction** shows that large models can memorise and regurgitate verbatim training examples, including secrets and personal data.[19] A fourth family, **model extraction/theft**, reconstructs a functional copy of a proprietary model through queries.[3]

> **DPO HINT**
>
> These attacks are the technical reason a model trained on personal data cannot automatically be assumed anonymous. If membership inference, inversion or extraction can recover personal data with non-negligible probability, the model itself may constitute processing of personal data — with direct consequences for lawful basis, data-subject rights and the anonymity analysis discussed in §6.2. Document the privacy attacks you tested for and their results.

3.4 Prompt injection and LLM-specific risks

Large-language-model applications introduce a distinctive class of vulnerabilities in which the boundary between trusted instructions and untrusted data is blurred. **Prompt injection** — where crafted input overrides the developer's instructions — has topped the OWASP LLM Top 10 for two consecutive editions.[8] Its indirect variant is especially dangerous: malicious instructions hidden in a web page, document or email are ingested by the model and executed on the user's behalf, a vector demonstrated in real data-exfiltration incidents against production assistants.[8] The 2025 OWASP list situates prompt injection alongside sensitive-information disclosure, supply-chain and data/model-poisoning risks, improper output handling, excessive agency, system-prompt leakage, vector and embedding weaknesses in RAG systems, misinformation, and unbounded consumption.[8]

> **CISO HINT**
>
> Design LLM applications on two principles: **treat all model output as untrusted** (validate and encode it before it reaches a browser, shell, database or downstream system), and **apply least privilege to agency** (constrain the tools and permissions a model can invoke, and require human approval for consequential or irreversible actions). Add input/output guardrails at an AI gateway and log prompts and tool calls for detection.

3.5 Agentic and multi-agent risks

As applications move from passive chat to autonomous agents that call tools, query databases and act, the attack surface shifts from the model's words to its actions. Excessive agency — granting an agent more tools, permissions or autonomy than its task requires — is among the most consequential new risks.[8] The OWASP Top 10 for Agentic Applications adds agent goal hijacking, tool misuse, and memory or context poisoning, while the NIST AML taxonomy now treats multi-agent "prompt worm" propagation and retrieval-store poisoning explicitly.[3][9] Governance frameworks are being extended in step, with agentic profiles and control matrices emerging to address delegation, runtime behaviour and accountability across agent chains.[26]

Table 2 — Attack taxonomy against AI systems, with cross-references.

Attack class	Mechanism	Primary references
Evasion / adversarial examples	Inference-time perturbation to force misclassification	NIST AML; Goodfellow et al.[3][20]
Data / model poisoning & backdoors	Training-time corruption; trigger-activated behaviour	NIST AML; OWASP LLM04; ENISA[3][8][11]

Attack class	Mechanism	Primary references
Supply-chain compromise	Poisoned hosted models, packages, coding-assistant configs	OWASP LLM03; ENISA[8][11]
Membership inference	Determine if a record was in training data	Shokri et al.[17]
Model inversion	Reconstruct sensitive inputs from outputs	Fredrikson et al.[18]
Training-data extraction	Regurgitation of memorised training examples	Carlini et al.[19]
Model extraction / theft	Query-based reconstruction of a model	NIST AML; OWASP[3][8]
Prompt injection (direct/indirect)	Untrusted input overrides instructions / actions	OWASP LLM01; NIST AML[3][8]
Sensitive-information disclosure	Model leaks PII, secrets or proprietary data	OWASP LLM02[8]
Excessive agency / agentic abuse	Over-permissioned tool use; goal hijacking; memory poisoning	OWASP LLM06 & Agentic; NIST[3][8][9]

4 AI for cyber offence: weaponisation

Beyond attacks on AI, AI is increasingly used as an offensive instrument. The clearest, best-evidenced effect is on social engineering: ENISA reports that AI-assisted phishing accounted for more than 80% of observed social-engineering activity by early 2025, with phishing responsible for roughly 60% of successful intrusions, augmented by AI-generated lures, deepfake voice and video, and purpose-built malicious LLMs such as WormGPT and FraudGPT.[11]

A second, faster-moving frontier is autonomy. Peer-reviewed and industry work through 2024–2025 established that language-model agents can autonomously exploit certain one-day (known) vulnerabilities, hack simple websites, and drive structured penetration-testing workflows, while multi-agent frameworks can reproduce vulnerabilities end-to-end from public CVE descriptions.[21][22][25] Vendor red-team programmes and initiatives such as Google's Project Naptime have shown frontier models assisting in vulnerability discovery with limited human input.[23] The literature frames the aggregate effect as cyber-threat inflation: AI lowers the skill and cost barriers to multi-stage intrusions, scaling and accelerating offensive activity rather than inventing wholly new attack physics.[25]

> **CISO HINT**
> Plan for a compressed exploitation timeline. ENISA notes attackers weaponising new vulnerabilities within days of disclosure,[11] and AI further shortens that window. Prioritise **rapid patching and exposure management**, **phishing-resistant MFA**, and detection tuned for AI-generated social engineering and deepfake-enabled fraud (out-of-band verification for high-risk requests). Assume commodity attackers now have expert-level tooling.

5 AI for cyber defence

The same capabilities strengthen defenders. Machine learning has long underpinned anomaly and intrusion detection, malware classification and spam filtering; the generative wave adds language-model assistance to the security-operations centre. Reported and researched applications include alert triage and summarisation, log and threat-intelligence analysis, detection-rule and query generation, vulnerability triage, phishing detection, and interactive honeypots that use language models to engage attackers.[24][25] Used well, these tools reduce analyst toil and compress mean-time-to-respond.

Two caveats recur in the literature. First, defensive AI inherits the vulnerabilities of §3 — a phishing-detection model can itself be targeted by prompt injection,[8] and an over-trusted assistant can hallucinate or be manipulated. Second, automation without oversight is itself a risk: confident, fluent output invites over-reliance. Defensive AI should therefore augment, not replace, expert judgement, with humans retaining authority over consequential actions.

> **BALANCE**
> The offence–defence balance is contested and evolving. The responsible posture is to adopt AI for defence deliberately — with evaluation, guardrails and human oversight — while assuming adversaries are adopting it faster and with fewer constraints.

6 The regulatory and data-protection dimension

Technical controls now sit inside a hardening legal frame. Two instruments dominate for European and Swiss organisations: the EU AI Act, which regulates AI systems by risk tier, and the GDPR, which governs any processing of personal data — including the personal data inside models.

6.1 The EU AI Act

The AI Act (Regulation (EU) 2024/1689) applies a risk-based approach, with the heaviest obligations falling on high-risk systems. Its Article 15 is the security keystone: high-risk AI systems must be designed to achieve an appropriate level of **accuracy, robustness and**

cybersecurity and to perform consistently across their lifecycle.[13] Crucially, the Article requires resilience against attempts by unauthorised third parties to alter a system's use, outputs or performance by exploiting vulnerabilities, and calls for technical measures to prevent, detect and respond to AI-specific attacks — explicitly including data poisoning and adversarial manipulation — with feedback-loop risks in continuously-learning systems to be mitigated.[13] Article 15 does not operate alone; it interlocks with the Act's obligations on risk management, data governance, documentation, logging and human oversight.

CISO HINT

If you operate high-risk AI, map Article 15 directly onto your control set: adversarial and poisoning-resilience testing, robustness and accuracy benchmarking with declared metrics, tamper-resistance, logging, and lifecycle monitoring for drift and feedback loops. ISO 42001 plus the NIST AML taxonomy give you a defensible way to evidence “appropriate” cybersecurity to conformity assessors.

6.2 GDPR and the EDPB Opinion 28/2024

On 17 December 2024 the European Data Protection Board adopted Opinion 28/2024, its authoritative interpretation of how the GDPR applies to personal data in AI models.[15] Three conclusions matter most for practitioners.

- **Anonymity is not automatic.** Not every model trained on personal data is anonymous; the threshold is high and must be assessed case-by-case. A model may be considered anonymous only where the probability of extracting personal data — directly or via queries — is negligible for every data subject.[15] The privacy attacks of §3.3 are precisely what makes this hard to demonstrate.
- **Legitimate interest can be a basis — conditionally.** Article 6(1)(f) may lawfully ground development and deployment, but only after a structured three-step legitimate-interest assessment (purpose, necessity, balancing) consistent with the EDPB's Article 6(1)(f) guidelines.[15][16]
- **Unlawful development can taint deployment.** If personal data is processed unlawfully during development, that unlawfulness may affect the lawfulness of the model's subsequent operation.[15]

DPO HINT

Build an evidence file per model: (1) a documented **legitimate-interest assessment** or other lawful basis; (2) a **DPIA** where processing is likely to be high-risk (large-scale, sensitive data, or systematic evaluation); (3) an **anonymity analysis** that records the extraction/inference tests performed and why residual risk is negligible; and (4) mechanisms for **data-subject rights** and transparency. A failure to document may itself indicate non-compliance.[15]

GDPR NOTE

Beyond Opinion 28/2024, keep the core principles in view: purpose limitation and data minimisation (Art. 5), lawful basis (Art. 6) and the special-category regime (Art. 9), transparency (Arts. 13–14), and the safeguards around solely automated decisions with legal or similarly significant effects (Art. 22). Where an AI system makes or materially shapes such decisions, ensure meaningful human involvement and the individual's right to contest.

6.3 Sectoral regimes and Switzerland

Depending on sector and market, further duties apply. NIS2 (Directive (EU) 2022/2555) raises baseline cybersecurity and incident-reporting obligations for essential and important entities, and DORA (Regulation (EU) 2022/2554) imposes ICT-risk and resilience requirements — including third-party risk — on financial entities; both are relevant when AI is part of the in-scope estate.[28] Swiss-based organisations remain subject to the revised Federal Act on Data Protection (revFADP/nDSG, in force since 1 September 2023) and, through market reach, frequently to the GDPR and the AI Act.[27]

7 Practitioner guidance

The preceding sections converge on a practical question: what should the two accountable roles actually do? This section distils the synthesis into role-specific programmes.

7.1 For the CISO

- **Govern.** Stand up AI governance (an AI register, an intake/approval gate, and a cross-functional committee), and adopt a management-system baseline (ISO 42001 mapped to the NIST AI RMF).[1][4]
- **Assess with a real threat model.** Populate risk assessments from the NIST AML taxonomy, OWASP Top 10 (LLM and Agentic) and MITRE ATLAS, not from generic checklists.[3][8][9][10]
- **Secure the pipeline.** Maintain an AI-BOM; verify model and dataset provenance and integrity; protect training and fine-tuning data against poisoning; harden and monitor infrastructure.[11]
- **Secure the application.** Treat model output as untrusted; enforce least-privilege agency and human approval for consequential actions; deploy input/output guardrails; log prompts and tool calls.[8]
- **Test adversarially.** Red-team models and agents for prompt injection, jailbreaks, poisoning and privacy leakage; benchmark robustness and accuracy — and, for high-risk systems, evidence this for Article 15.[3][13]
- **Detect and respond.** Extend the SOC to AI-specific telemetry and incident scenarios; prepare for AI-accelerated phishing, deepfakes and rapid exploitation.[11]
- **Manage third parties.** Assess AI vendors' security, data handling and training-data practices; contract for audit rights, incident notification and data return.

7.2 For the Data Protection Officer

- **Establish lawful basis.** Document the basis for each processing activity; where relying on legitimate interest, complete and retain the three-step assessment.[15][16]
- **Run DPIAs.** Complete a DPIA wherever AI processing is likely to be high-risk, and integrate it with the AI Act's fundamental-rights impact assessment where both apply (see §7.3).
- **Interrogate anonymity claims.** Require an evidenced anonymity analysis before treating a model or its outputs as anonymous, informed by membership-inference, inversion and extraction testing.[15][17][18][19]
- **Govern training data.** Confirm provenance, lawful basis and minimisation for training, fine-tuning and retrieval data; control third-party data and the use of organisational data to train external models.
- **Uphold rights and transparency.** Provide the required information to individuals, honour access, objection and erasure, and safeguard solely-automated decisions under Article 22.
- **Prepare for incidents.** Treat model data leakage (e.g. via extraction or prompt injection) as a potential personal-data breach and align with breach-notification duties.

7.3 Where the AI Act and GDPR intersect

The two regimes overlap but do not coincide, and the DPO and CISO should operate them jointly rather than in parallel silos. Table 3 highlights the practical touchpoints.

Table 3 — Practical intersections of the EU AI Act and the GDPR.

Theme	EU AI Act	GDPR
Impact assessment	Fundamental-rights impact assessment (FRIA) for certain deployers	DPIA for high-risk processing — run them together where both apply
Security	Art. 15 accuracy, robustness, cybersecurity for high-risk AI	Art. 32 security of processing (appropriate technical/organisational measures)
Human oversight	Art. 14 human oversight of high-risk AI	Art. 22 safeguards for solely-automated decisions
Transparency	Disclosure of AI interaction; instructions for use	Arts. 13–14 information duties to data subjects
Data governance	Art. 10 data and data-governance for high-risk AI	Art. 5 principles; lawful basis; minimisation

GDPR NOTE

Where a system is both high-risk under the AI Act and processes personal data, combine the FRIA and DPIA into a single, coordinated assessment. This avoids duplicated effort, produces one coherent evidence file, and ensures the CISO's Article 15 security measures are recognised as part of the DPO's Article 32 assurance.

8 Open challenges and research directions

- **Robust defences remain unsolved.** No general defence exists against adversarial examples, prompt injection or poisoning; each is an active arms race.[3][8]
- **Agentic accountability.** Delegation chains, tool use and runtime behaviour outpace current governance; agentic profiles and control matrices are early-stage.[9][26]
- **Measuring anonymity and memorisation.** Standardised, defensible methods for demonstrating negligible extraction risk are needed to operationalise Opinion 28/2024.[15][19]
- **Benchmarking offence and defence.** Realistic, comparable evaluations of AI cyber capability are still maturing, complicating risk quantification.[25]
- **Regulatory operationalisation.** Article 15's "appropriate" levels await harmonised benchmarks and standards; the case-by-case posture of Opinion 28/2024 leaves interpretive room across authorities.[13][15]

9 Conclusion

Cybersecurity and AI can no longer be treated as separate disciplines. AI systems must be defended against a distinctive and fast-growing catalogue of attacks; AI is simultaneously reshaping the offence–defence balance; and a hardening regulatory frame — the EU AI Act's security obligations and the GDPR as interpreted by the EDPB — now converts much of this into legal duty. The organisations that navigate this well will be those that unite three capabilities that have historically sat apart: security engineering, AI governance, and data protection. In practice that means the CISO and the DPO working from a shared threat model and a shared evidence base. This paper has offered that shared map, drawn from the field's leading standards and literature, and translated it into the concrete actions each role can take now.

— References

Sources are cited in the text by bracketed number. Titles are given for identification; consult the original publications for authoritative text. This paper is a synthesis and does not reproduce source material.

1. NIST. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023.
2. NIST. *Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile*, NIST AI 600-1, 2024.
3. Vassilev, A., Oprea, A., et al. *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2e2025, 2025.
4. ISO/IEC 42001:2023, *Information technology — Artificial intelligence — Management system*.
5. ISO/IEC 27001:2022, *Information security management systems — Requirements*.
6. ISO/IEC 23894:2023, *Artificial intelligence — Guidance on risk management*.
7. ISO/IEC 27090 (AI security guidance) and ISO/IEC 27091 (AI privacy), under development.
8. OWASP Foundation. *OWASP Top 10 for Large Language Model Applications (2025)*.
9. OWASP Foundation. *OWASP Top 10 for Agentic Applications (2025)*.
10. MITRE. *ATLAS — Adversarial Threat Landscape for Artificial-Intelligence Systems*.
11. ENISA. *Threat Landscape 2025* (reporting period Jul 2024 – Jun 2025), 2025.
12. ENISA. *Artificial Intelligence Cybersecurity Challenges — AI Threat Landscape*.
13. Regulation (EU) 2024/1689 (*EU Artificial Intelligence Act*), esp. Article 15.
14. Regulation (EU) 2016/679 (*General Data Protection Regulation, GDPR*).
15. European Data Protection Board. *Opinion 28/2024 on certain data protection aspects related to the processing of personal data in the context of AI models*, 17 Dec 2024.
16. European Data Protection Board. *Guidelines 1/2024 on processing based on Article 6(1)(f) GDPR (legitimate interests)*.
17. Shokri, R., Stronati, M., Song, C., Shmatikov, V. *Membership Inference Attacks Against Machine Learning Models*, IEEE S&P, 2017.
18. Fredrikson, M., Jha, S., Ristenpart, T. *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, ACM CCS, 2015.
19. Carlini, N., et al. *Extracting Training Data from Large Language Models*, USENIX Security, 2021 (and related extraction work).
20. Goodfellow, I., Shlens, J., Szegedy, C. *Explaining and Harnessing Adversarial Examples*, ICLR, 2015.
21. Fang, R., et al. *LLM Agents can Autonomously Exploit One-Day Vulnerabilities / Autonomously Hack Websites*, 2024.
22. Deng, G., et al. *PentestGPT: Evaluating and Harnessing LLMs for Automated Penetration Testing*, USENIX Security, 2024.
23. Glazunov, S., Brand, M. *Project Naptime: Evaluating Offensive Security Capabilities of Large Language Models*, Google Project Zero, 2024.
24. *Security Concerns for Large Language Models: A Survey*, 2025.
25. *LLM-based Agents in Autonomous Cyberattacks* (“Forewarned is Forearmed”), 2025.
26. Cloud Security Alliance. *Agentic AI — NIST AI RMF Agentic Profile* and *AI Controls Matrix (AICM)*, 2025.
27. Swiss Confederation. *Revised Federal Act on Data Protection (revFADP / nDSG)*, in force 1 Sep 2023.
28. Directive (EU) 2022/2555 (*NIS2*); Regulation (EU) 2022/2554 (*DORA*).

This document is a research synthesis for informational purposes and does not constitute legal advice. Regulatory interpretations evolve; verify current requirements with qualified counsel and the primary sources.