

PERSPECTIVE · OCTOBER 2026

Understanding *ISO 27090* (AI Security)

A briefing for boards and senior leaders, October 2026

Executive Summary

- **The timing is right.** ISO/IEC 27090, the international guidance on attacks against AI systems, was published on 14 October 2026, just as concern about AI security has turned into something close to paranoia in boardrooms.
- **It explains how AI systems are attacked and what to do about it.** In 56 pages, the committee behind ISO 27001 describes nine ways AI systems are compromised and, for each, what fails, how to detect the attack and how to limit it.
- **It is guidance and should be used as guidance.** It is not a checklist, and there is no ISO 27090 certificate to frame and hang in reception.
- **It does not deliver EU AI Act compliance.** An ISO 27090 review gives no presumption of conformity; that route runs through European standards still being finalised.
- **It stands on other people's work.** It builds on ISO 27001 and 27002, cites ETSI's baseline AI security standard and absorbs around 70 pages from OWASP's free AI Exchange. Much of its substance is available at no cost.
- **Its greatest value is gap analysis.** Firms can hold their existing frameworks and chosen standards against its structure and see where AI risk falls through.
- **Three areas need work beyond it.** AI agents get a single heading, attackers using AI are out of scope, and the portions ISO publishes free do not say how to prove that a model has not been tampered with.

AI security has become a board question, and ISO 27090 answers half of it

AI has moved from pilots into systems that approve payments, triage claims, answer customers and, increasingly, act on their own. These systems fail under attack in ways conventional IT does not. An attacker does not need to break in to cause harm: altering the data a model learns from, asking a model enough questions to copy it, or hiding instructions in a document the AI will read are each enough. In October 2025, research by Anthropic, the UK AI Security Institute and the Alan Turing Institute found that 250 poisoned documents were enough to plant a hidden trigger in language models of every size tested, up to 13 billion parameters. For the largest model, those documents made up 0.00016% of its training data.

Supervisors are moving. The Monetary Authority of Singapore (MAS) finalised its guidelines on AI risk management on 7 October 2026, effective a year later. In the EU, the AI Act requires high-risk AI systems to be accurate, robust and secure, and a political agreement in June 2026 sets those obligations to apply from December 2027 and August 2028.

Boards are asking two distinct questions: can attackers use AI against us, and can attackers turn our own AI against us? ISO 27090 addresses the second. Please do not treat the new ISO standard as a complete solution.

Four years of drafting produced a method different to a rulebook

Work began in September 2022 within ISO/IEC JTC 1/SC 27, the committee responsible for ISO 27001. A draft passed its international vote in July 2025 and grew from 46 to 56 pages before publication. The title changed on the way: “security threats and failures” became “security threats and compromises”. The change reflects the focus. The standard deals with deliberate attack and treats failure as the consequence of an attack, with accidental malfunction left largely to other standards.

The standard has three parts:

- **Applying existing security discipline to AI:** governance, zero trust, secure engineering, supply chain management, threat modelling and red teaming. ISO 27001 and 27002 are the foundation, with AI-specific measures added on top.
- **Nine ways AI systems are compromised:** each written to the same template of what the attack is, what fails as a result, how to detect it and how to mitigate it.
- **How defences behave over time:** conflicts between controls, decay, logging, protection of development environments, filtering of malicious inputs and limits on model use.

Annexes map attacks to the stages of the AI life cycle and to the assets at risk, describe AI versions of conventional attacks (denial of service, supply chain compromise, backdoors, and theft of models and data) and cover federated learning, where a model is trained across many devices without pooling the data.

COMPROMISE	WHAT HAPPENS	BUSINESS CONSEQUENCE
Data poisoning	An attacker alters the data a model learns from, including data used for retraining	Wrong decisions at scale; hidden triggers that activate on cue
Evasion	Inputs are altered slightly, often invisibly to humans, so the model misjudges them	Fraud, spam or malware passes the AI filter
Membership inference	An attacker works out whether a person's record was used in training	A privacy breach without any data leaving the firm

COMPROMISE	WHAT HAPPENS	BUSINESS CONSEQUENCE
Model exfiltration	An attacker copies a model's behaviour by asking it enough questions	Loss of intellectual property through legitimate-looking use
Model inversion	An attacker reconstructs training data from the model's answers	Exposure of personal or confidential data
Direct model poisoning ¹	The model itself is tampered with through code, a supplier or the pipeline	Malicious behaviour that survives later fine-tuning
Sensitive model output	The model reveals training data, other users' prompts or internal instructions	A data breach through normal use
Prompt injection	Instructions are hidden in user input or in documents, emails and web pages the AI reads	The AI acts for the attacker, leaking data or misusing tools
Output containing injection attacks	The AI produces content that attacks the systems consuming it	Conventional compromise of downstream systems

The template gives leaders four questions to ask of any AI system: what can go wrong, what would fail, would we notice, and what stops it?

The standard's most valuable advice concerns how defences behave

Three points rarely appear in summaries of ISO 27090, and each matters for investment decisions.

Defences conflict. A control that blocks one attack can open the door to another. The standard's own example: training a model on adversarial examples, a common hardening technique, made it markedly more susceptible to patch attacks, in which a visible sticker or pattern fools image recognition. Controls should be assessed together before deployment, and a firm should record how it resolved each trade-off. The standard's own headings contain one: minimising published technical detail about a model pulls against transparency and explainability.

Defences decay. Models drift, and controls that depend on training lose effect as the model changes. Every retraining or model update changes the security posture and needs revalidation before the system returns to service.

Detection comes first, and for some attacks it does not yet exist. The standard treats mitigation without effective detection as of limited use. It also states plainly that detecting

¹ We had predicted poisoning as an attack vector. <https://santoshpandit.com/2026-cti-forecast-part-2.pdf> (page 17) for those interested.

poisoned data or models remains an open research field, that automated tools struggle to separate malicious data from harmless outliers, and that unusual inputs are not necessarily attacks. Where detection is immature, the money belongs in prevention: knowing and verifying where data and models come from.

One further detail matters for testing. The standard classes attackers by what they know about the system: zero, partial or perfect knowledge. These map directly onto black-box, grey-box and white-box testing. A board commissioning AI testing can fix the knowledge level in the contract. Perfect-knowledge testing is the hardest case, and the one that a well-resourced attacker or an insider represents.

ISO 27090 is not a checklist, a certificate or a compliance route

It is not a checklist. The standard describes its examples as illustrative and expects each organisation to relate threats to its own systems according to risk. Ticking off nine threats says nothing about whether the right ones were treated well.

It is not a certificate. It contains almost no requirements: the final draft used “shall” only twice, to require ISO 27002 measures and consideration of AI-specific attacks across the life cycle. No certification scheme exists, and guidance of this kind cannot support one. Organisations certify to ISO 27001 or ISO 42001 and use ISO 27090 to inform the risk assessment beneath them. A supplier claiming to be “ISO 27090 certified” is probably selling ISO-washed snake oil.

It is not EU AI Act compliance. Presumption of conformity under the AI Act comes only from harmonised standards cited in the Official Journal of the EU. For AI cybersecurity the candidate is prEN 18282, whose public enquiry closed on 30 July 2026, with ETSI EN 304 223 alongside it. No published mapping links either to ISO 27090. Assuming that an ISO 27090 review makes a firm AI Act compliant is simply wrong.

It is not yet a supervisory expectation. The UK government listed the draft as work in progress in its January 2025 guide to the AI Cyber Security Code of Practice. The MAS guidelines of 7 October 2026 cite ISO/IEC 22989 and the NIST AI Risk Management Framework and do not mention ISO 27090; the Prudential Regulation Authority’s model risk statement SS1/23 predates it. Firms that adopt its structure now will be ready if supervisors adopt it later.

As a former regulator, I continue to hold a view that regulators should not prescribe specific standards such as ISO 27001, 42001, or 27090 as solutions.

ISO 27090 stands on work, much of which is free

ISO 27090 builds on the ISO 27000 family and on two ISO foundations for AI: ISO/IEC 22989 for terminology and life-cycle stages, which ISO offers free of charge, and ISO/IEC 23053 for machine learning systems. The final draft cites ETSI EN 304 223, Europe's free baseline standard for AI security. Its threat catalogue lines up closely with NIST's free taxonomy of adversarial machine learning and with MITRE ATLAS, the public knowledge base of attacks on AI.

Its largest debt is to OWASP. The OWASP AI Exchange, a free and openly licensed body of AI security guidance, states that it contributed around 70 pages through a formal liaison with the ISO committee. Our own mapping bears this out: the AI Exchange tracks ISO 27090's guidance on limiting data and governing model behaviour heading for heading, and uses the same names for several attacks.

The implication for leaders is practical. The paid text adds international consensus, a stable vocabulary that auditors and procurement teams can cite, and detail on the limits of each defence. The substance needed to start work is free today. In the UK, British Standards Online, which several public libraries offer free on site, lists the UK edition, BS ISO/IEC 27090:2026, for 31 October 2026.

Gap analysis is where ISO 27090 earns its keep

Most firms already run several frameworks: ISO 27001, often ISO 42001, the NIST AI Risk Management Framework, OWASP's Top 10 lists, and sector rules such as the EU's Digital Operational Resilience Act (DORA) or SS1/23. ISO 27090 gives them a single lens to hold all of these against. We recommend five steps:

- **Inventory every AI system**, including third-party services and AI embedded in ordinary software, with its data sources, models, connections and owners.
- **Map each system to the nine compromises and the conventional attacks in the annexes**, stage by stage, and set aside those that do not apply.
- **Test current controls against the four-part template**. For each relevant compromise, the firm should be able to say what would fail, show that it would detect the attack and point to a control that limits it. Detection gaps matter most, because the standard treats mitigation without detection as of limited use.
- **Review controls as a set**. Record conflicts and trade-offs, and require revalidation after every retraining or model change.
- **Extend beyond the standard** where it is thin, as set out below.

Mapping shows where familiar frameworks stop. OWASP's analysis finds that ISO 27002, the control catalogue behind most ISO 27001 certificates, has no control for throttling the queries that copy a model, and that ISO standards do not yet set controls for minimising AI training data or limiting its retention. An ISO 27001 certificate will not reveal these gaps.

One practical warning. National standards bodies treat loading standards text into AI tools as a breach of licence: Standards New Zealand stated in February 2026 that uploading standards to ChatGPT or similar tools is strictly prohibited. Firms planning to automate their gap analysis should check their licence before giving the standard to an AI assistant.

Three gaps firms must fill themselves

AI agents. Only one heading addresses agents directly: least privilege for AI systems and agents. Agents that read email, call tools and take actions create risks the standard barely touches, including misuse of tools, privilege escalation across chains of agents and actions no human approved. OWASP's Top 10 for Agentic Applications (2026), joint guidance from national cyber agencies including the UK National Cyber Security Centre (May 2026) and Singapore's agentic addendum (June 2026) cover much of this ground.

Attackers using AI. ISO 27090 protects AI systems; it does not address what AI does in an attacker's hands. That threat needs its own programme: faster patching, AI-assisted testing of the firm's own estate and recovery planning.

Proof that the model has not been tampered with. The standard asks firms to know where their AI systems and data come from, but the portions ISO publishes free do not say how to prove it. A record of origin that anyone with access can edit proves little. Boards should expect their teams to show that the model running in production is the one that was approved, and that its training data has not been altered since. The technology to do this exists and should be planned now, including protection that will withstand future quantum computers, because models and their records are kept for years.

Questions for the Board this quarter

- Do we have a complete inventory of the AI systems we build, buy and use, including AI embedded in suppliers' software?
- For our most important AI systems, which of the nine compromises apply, and would we detect each one?
- Who revalidates security after a model is retrained or replaced, and when did that last happen?

- Have we tested our AI with testers given full knowledge of the system, and what did they fix?
- What can our AI agents do without human approval, and who set those limits?
- Can we prove that the model running in production is the one we approved, and that nobody has altered it since?
- Are we relying on any framework, certificate or review, ISO 27090 included, as evidence of EU AI Act compliance?

ISO 27090 gives organisations a shared language for attacks on AI and a disciplined way to think about defending against them. Used as a lens for gap analysis, it shows where existing frameworks stop.

We will be happy to bring the benefit of our experience and regulatory expertise to boards and security teams applying ISO 27090. If you are in the UK, you can of course buy the standard from BSI for GBP 322. Or you can invite us for a coffee discussion that costs you GBP 3.22. Or do both!

Santosh Pandit · Board and CxO Advisor · Founder, BeatQuantum Limited

BeatQuantum.com | SantoshPandit.com | contact@santoshpandit.com

References

1. ISO, ISO/IEC 27090:2026 Cybersecurity - Artificial intelligence (AI) - Addressing security threats and compromises to AI systems, catalogue entry and free sample. <https://www.iso.org/standard/27090>
2. ISO, ISO/IEC 27090 explained. <https://www.iso.org/insights/ISO-27090-explained>
3. Anthropic, UK AI Security Institute and Alan Turing Institute, A small number of samples can poison LLMs of any size, 9 October 2025. <https://www.anthropic.com/research/small-samples-poison>
4. Monetary Authority of Singapore, Guidelines on Artificial Intelligence Risk Management, 7 October 2026. <https://www.mas.gov.sg/-/media/mas-media-library/regulation/guidelines/bd/guidelines-on-artificial-intelligence-risk-management-for-financial-institutions/guidelines-on-artificial-intelligence-risk-management.pdf>
5. Sidley Austin, EU lawmakers reach provisional agreement to delay key EU AI Act obligations, 22 June 2026. <https://datamatters.sidley.com/2026/06/22/eu-lawmakers-reach-provisional-agreement-to-delay-key-eu-ai-act-obligations/>
6. ISO, ISO/IEC 27090 project history, committee page. <https://committee.iso.org/standard/56581.html>
7. Secuvis, ISO/IEC 27090: security standard for AI systems (analysis of the final draft), 13 August 2026. <https://www.cyber-regulierung.de/iso-27090/>
8. Secuvi, ISO/IEC 27090: the security standard for AI systems. <https://www1.secui.com/blog/compliance/iso-iec-27090-security-standard-for-ai-systems/>
9. Gary Hinson, ISO/IEC 27090 commentary, ISO27001security.com. <https://www.iso27001security.com/html/27090>
10. Adam Leon Smith, Which standard to use for the cybersecurity of AI, December 2025. <https://adamleonsmith.substack.com/p/which-standard-to-use-for-the-cybersecurity>

11. Nannini, Leon Smith and others, analysis of European AI cybersecurity standards, arXiv 2604.04604, April 2026. <https://arxiv.org/html/2604.04604v1>
12. CEN-CENELEC JTC 21, Significant milestone for European AI standardisation (prEN 18282 enquiry). <https://jtc21.eu/significant-milestone-for-european-ai-standardization/>
13. ETSI EN 304 223 V2.1.1, Securing Artificial Intelligence: Baseline Cyber Security Requirements for AI Models and Systems, December 2025. https://www.etsi.org/deliver/etsi_en/304200_304299/304223/02.01.01_60/en_304223v020101p.pdf
14. Department for Science, Innovation and Technology, Implementation Guide for the AI Cyber Security Code of Practice, January 2025. https://assets.publishing.service.gov.uk/media/679cae441d14e76535afb630/Implementation_Guide_for_the_AI_Cyber_Security_Code_of_Practice.pdf
15. Prudential Regulation Authority, SS1/23 Model risk management principles for banks. <https://www.bankofengland.co.uk/-/media/boe/files/prudential-regulation/supervisory-statement/2023/ss123.pdf>
16. ISO, ISO/IEC 22989 Artificial intelligence concepts and terminology (free). <https://committee.iso.org/standard/74296.html>
17. NIST AI 100-2 E2025, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, March 2025. <https://csrc.nist.gov/pubs/ai/100/2/e2025/final>
18. MITRE ATLAS, Adversarial Threat Landscape for Artificial-Intelligence Systems. <https://atlas.mitre.org>
19. OWASP AI Exchange, AI security overview. https://owaspai.org/docs/ai_security_overview/
20. BSI Knowledge, BS ISO/IEC 27090:2026. <https://knowledge.bsigroup.com/products/cybersecurity-artificial-intelligence-ai-addressing-security-threats-and-compromises-to-ai-systems>
21. Standards New Zealand, Using standards content in digital and AI tools: what you need to know, 27 February 2026. <https://www.standards.govt.nz/news-and-updates/using-standards-content-in-digital-and-ai-tools-what-you-need-to-know>
22. OWASP, Top 10 for Agentic Applications for 2026. <https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/>
23. National cyber agencies including the UK NCSC, Careful adoption of agentic AI services, May 2026. <https://ncsc.govt.nz/protect-your-organisation/careful-adoption-of-agentic-ai-services/>
24. Cyber Security Agency of Singapore, Addendum on Securing AI Systems (agentic AI), June 2026. <https://www.csa.gov.sg/resources/publications/addendum-on-securing-ai-systems/>