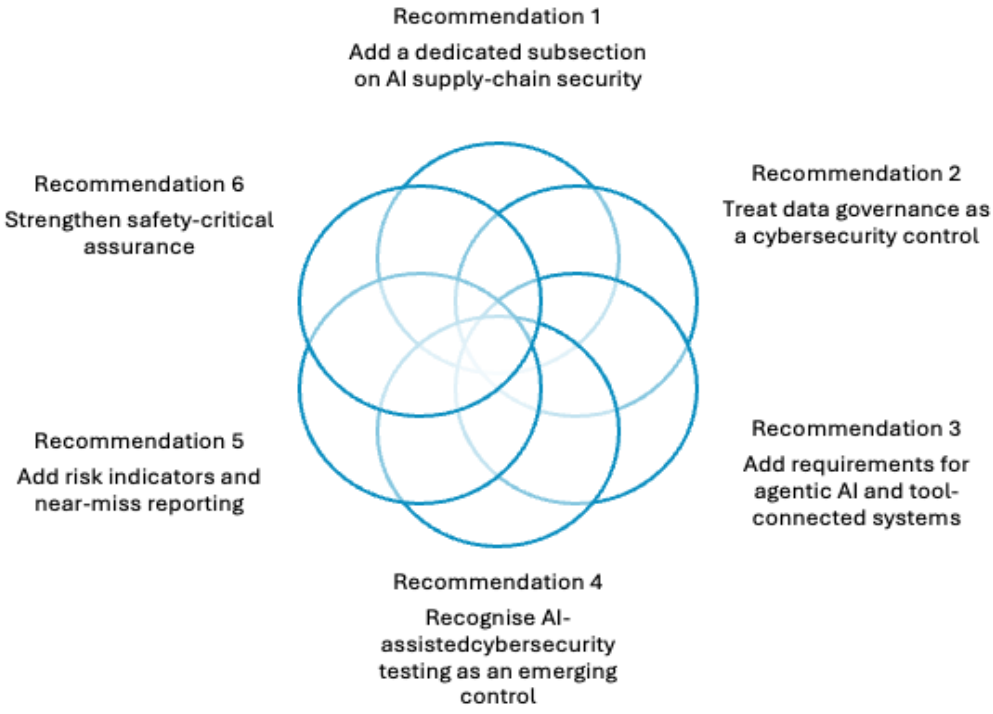
Reflections on Draft prEN 18282: Cybersecurity Specifications for AI Systems

July 2026

The draft standard on Cybersecurity Specifications for AI Systems is a timely and valuable contribution to the European AI governance landscape.[i] It correctly recognises that cybersecurity for AI systems is not limited to conventional ICT security. AI systems introduce additional attack surfaces through training data, validation data, inference data, model artefacts, prompts, configurations, deployment components, APIs, software agents, third-party models, and operational environments. Recent research similarly shows that AI security risks arise across the AI lifecycle, including data collection, model training, inference, deployment, and integration with downstream applications [1, 2].

The draft standard also rightly places cybersecurity of AI systems within the broader lifecycle of an AI system, requiring providers to determine relevant circumstances, identify vulnerabilities and threats specific to AI systems, assess cybersecurity risks posed to AI systems, select measures, perform testing, and document the results.[ii]

We welcome the draft standard and recommend that it be strengthened in six areas, detailed in the recommendations section below:



Scope and Guiding Principles

Our consultation in response to the standard focused on the cybersecurity protection of AI systems within the scope of CEN/CENELEC prEN 18282. It did not seek to expand the standard into a broader framework for general AI risk management, societal risk, or sector-wide governance. Those issues are more appropriately addressed in related instruments, including CEN/CENELEC prEN 18228:2026 on AI Risk Management. The value of prEN 18282 lies in providing a clear and technically useful standard for the cybersecurity and AI communities, enabling providers/ deployers of AI systems to protect AI systems against evolving cyber threats while maintaining consistency with the EU AI Act standardisation framework.

Our response is guided by **four principles**:
First, AI supply-chain security should explicitly cover the full range of components that can affect the confidentiality, integrity, availability, or intended behaviour of an AI system. This includes data, models, tools, connectors, externally hosted services, software dependencies, retrieval sources, prompts, model updates, and the associated provenance, integrity, update, assurance, and rollback arrangements.

Second, retrieval-augmented and agentic systems require explicit controls for trust boundaries, permissions, tool use, external content, and connected services. Where an AI system can access files, APIs, repositories, browsers, communication channels, ticketing systems, cloud services, code execution environments, or deployment pipelines, cybersecurity assurance should address not only the model but also the operational autonomy (agents and self-healing systems) granted to the system.

Third, cybersecurity testing should be reproducible, subject to appropriate human validation, and responsive to changes in the threat environment. Testing should not be triggered only by internal design changes. It should also be reconsidered when there are new relevant threats, supplier incidents, vulnerabilities, changes to external dependencies, changes to data trust sources, or changes to tool permissions or connected services.

Fourth, where an AI system is deployed, the standard should encourage assurance that the system can enter a safe or controlled state, recover from a verified baseline, provide sufficient traceability of what occurred, and ensure that any residual cybersecurity risk is accepted through an accountable decision. This does not require prEN 18282 to be adopted as a general safety standard. Rather, it ensures that cybersecurity measures remain meaningful in deployments where compromise of the AI system may affect safety, security, operational continuity, or critical services.

The recommendations in this response are therefore intended as practical guidance rather than additional obligations. In particular, where appropriate the preferred formulation is generally “should” rather than “shall”, unless a mandatory requirement is clearly necessary for

consistency with the structure of the draft standard. This is important for implementability, especially for small and medium-sized organisations, and supports proportionate adoption without weakening the need for sound cybersecurity governance.

Analysis

The draft standard is particularly strong in recognising attacks specific to AI systems such as data poisoning, model poisoning, adversarial manipulation, confidentiality attacks, prompt injection, model flaws, and attacks through supply-chain components. It also usefully frames cybersecurity measures around the sequence of identify, protect, detect, respond, resolve, and control, which is appropriate for systems where prevention alone cannot eliminate risks to AI systems.

However, the draft standard should be strengthened in one important respect: AI supply-chain security should be treated as a first-order governance and cybersecurity concern, not merely as one element among several technical vulnerabilities. The current evolution of AI systems shows that the AI supply chain includes not only code, libraries, infrastructure, and APIs, but also datasets, model weights, pre-trained components, fine-tuning data, retrieval corpora, vector stores, prompts, system instructions, hosted model services, plugins, agent skills, tool connectors, and model-generated security outputs.

Research on large-language-model supply chains describes this ecosystem as spanning infrastructure, foundation models, downstream applications, data, model artefacts, and operational dependencies, while recent work related to AI systems highlights malicious AI models as an emerging software supply-chain concern [3, 6].

This is consistent with recent trends in cybersecurity risks, including the AI supply-chain security report, which highlights emerging concerns about model supply-chain compromise, malicious agent capabilities, hosted model security, prompt injection, model poisoning, MCP/tool security, and AI data exfiltration.

The draft standard, prEN 18282, is already aligned with a proportionate, risk-based approach. It acknowledges that cybersecurity risks specific to AI systems cannot be fully eliminated and that controls should be interpreted in light of their intended purpose, operational context, foreseeable misuse, and state of the art. This is appropriate. Nevertheless, the current threat environment has evolved quickly.

Some Frontier AI systems for cybersecurity and vulnerability research, their early testing processes indicate that such AI tools can materially assist vulnerability discovery, validation, patch generation, and remediation. This has been highlighted by the comprehensive EU Action Plan on Cybersecurity and Artificial Intelligence adopted on 7 July 2026. At the same time, the same capabilities may reduce the time and expertise required to identify and exploit

weaknesses, reinforcing the need for robust governance, human validation, controlled access, and responsible disclosure practices.

Previous work on LLMs for code security and vulnerability repair support the dual-use concern: such systems can assist security hardening and vulnerability detection, but their outputs require careful validation, reproducibility, and human review [10, 11].

We welcome the draft standard while recommending that it be strengthened in four areas:

1. AI supply-chain security should be made explicit and systematic.
2. Data should be treated as an asset to be protected, not merely as a quality or performance input.
3. Agentic and frontier AI capabilities should be addressed as both defensive tools and sources of accelerated cyber risk.
4. Safety-critical systems, including critical infrastructure and cyber-physical critical infrastructure, should require a stricter assurance overlay.

Recommendations

Recommendation 1: Add a dedicated subsection on AI supply-chain security

Providers should identify, document, and maintain an AI supply-chain security register for AI system components that can affect the confidentiality, integrity, availability, intended behaviour, or safety of the AI system. The register should include, where relevant, datasets, pre-trained models, model weights, fine-tuning data, validation and benchmark data, retrieval sources, vector stores, system prompts, embedded instructions, third-party APIs, hosted model services, plugins, agent tools, software dependencies, deployment components, and model update mechanisms.

The provider should assess the provenance, licensing, integrity, access control, update history, known vulnerabilities, security posture, and rollback options for such components. Where third-party or externally hosted components are used, the provider should document relevant contractual, technical, and operational dependencies, including data retention, model update, logging, confidentiality, and incident notification arrangements.

This reflects the growing recognition that AI supply-chain assurance must address both conventional software dependencies and AI-specific dependencies, including model provenance, supplied artefacts, data sources, hosted services, and downstream integrations [3, 6].

A recognised baseline cybersecurity certification or independently assessed assurance scheme may provide useful evidence of general cyber hygiene. However, such certification should not be treated as sufficient in itself for AI supply chain assurance. AI-related dependencies introduce additional concerns, including model provenance, integrity of model packages and updates, data provenance, retraining or fine-tuning practices, security of hosted services, tool permissions, connector behaviour, vulnerability disclosure arrangements, and the ability to suspend, quarantine, or roll back a compromised component.

The standard should therefore encourage providers to apply supply-chain requirements in a manner proportionate to the supplier's role, access, and the consequences of compromise. Evidence may include independent certification, security assessment reports, contractual commitments, technical testing, vulnerability management arrangements, incident notification obligations, and documented provenance and rollback controls. This approach would preserve technological neutrality across Europe while recognising that baseline organisational cybersecurity assurance and AI-specific supply-chain assurance serve complementary purposes.

Recommendation 2: Treat data governance as a cybersecurity control

Data used by or connected to an AI system shall be treated as a cybersecurity-relevant asset.

Previous work on poisoning attacks, shows that compromised training or fine-tuning data can directly alter model behaviour, and by privacy research showing that models may expose sensitive information about training data through inference or extraction attacks [2, 4, 5].

Providers shall implement appropriate controls for data provenance, lineage, integrity, access control, versioning, retention, licensing, and modification history. These controls shall apply to training data, validation data, testing data, inference data, fine-tuning data, retrieval sources, synthetic data, benchmark data, and operational feedback data, where relevant.

For systems that use retrieval-augmented generation or external knowledge sources, providers should assess the risk of retrieval poisoning, indirect prompt injection, source manipulation, unauthorised disclosure, and the propagation of untrusted content into model outputs or downstream actions. Research on indirect prompt injection shows that externally supplied content can manipulate LLM-integrated applications by blurring the boundary between data and instructions, which is particularly relevant for retrieval-augmented systems [7, 9].

Recommendation 3: Add requirements for agentic AI and tool-connected systems

Where an AI system can access tools, APIs, files, repositories, browsers, communication channels, ticketing systems, cloud services, code execution environments, or deployment

pipelines, the provider of the AI system should identify the permissions granted to the AI system and assess the cybersecurity risks to the AI system arising from such access. Studies of tool-integrated LLM agents show that indirect prompt injection can lead to unintended tool use, harmful actions, or data exfiltration, supporting the need for explicit controls over permissions, trust boundaries, and connected services [8].

The provider should apply least-privilege access, sandboxing, secrets isolation, tool-call logging, human approval for high-impact actions, and rollback mechanisms. The AI system should not autonomously modify safety-relevant data, production code, model parameters, access-control settings, deployment configurations, or operational workflows unless such actions are explicitly authorised, logged, tested, and subject to appropriate human oversight.

The deployer should configure, operate and monitor the AI system in accordance with those instructions and the relevant deployment context. This should include, where relevant, assigning and periodically reviewing access permissions; approving data sources and tool connections; protecting credentials and secrets; enabling and reviewing logs; maintaining human oversight; applying updates and mitigations; and reporting relevant vulnerabilities, incidents, near misses, or material configuration changes to the provider.

Neither the provider nor the deployer should permit the AI system to autonomously modify safety-relevant data, production code, model parameters, access-control settings, deployment configurations, or operational workflows unless such actions are explicitly authorised, logged, tested, subject to appropriate human oversight, and supported by documented recovery arrangements.

Provider responsibility	Deployer responsibility
Design the system so that least privilege, sandboxing, logging, approval gates and rollback are technically possible.	Configure and operate those controls correctly in the actual deployment.
Identify foreseeable risks associated with tool access and provide secure defaults, instructions, and limitations.	Assign users, roles, credentials, data sources and tool permissions according to the approved use case.
Define tested operating conditions and prohibited or unsupported configurations.	Do not deploy outside those conditions without reassessment.

Provide capability for audit logging, isolation, revocation, rollback and incident support.	Enable logging, protect logs, monitor alerts, revoke access and activate incident procedures.
Test the AI system and its supported integrations.	Test the deployed configuration, including local tools, repositories, data sources and workflows.
Issue security updates, vulnerability notices and mitigations.	Apply updates, assess local impact, notify the provider of incidents or changed conditions, and follow mitigation instructions.

Table 1: Allocation of Cybersecurity Responsibilities Between AI System Providers and Deployers for Agentic and Tool-Connected AI Systems

The division of responsibility is important because the provider often controls the design of safeguards and supported integrations, while the deployer controls the live configuration, local data sources, credentials, permissions, monitoring, and operational response.

Recommendation 4: Recognise AI-assisted cybersecurity testing as an emerging control

Where AI systems are used to support cybersecurity testing, vulnerability discovery, exploitability analysis, patch generation, code review, or red-team activities, the provider shall ensure that AI-generated findings are validated through appropriate technical and human review. AI-generated findings shall include sufficient evidence to support reproducibility, triage, severity assessment, and mitigation. Peer-reviewed research on LLMs for code security and vulnerability repair not only supports the potential of these tools for security hardening, but also shows why reproducibility, triage, patch validation, and human review remain necessary [10, 11].

Such evidence should include, where relevant, affected components, vulnerability hypothesis, test case or proof of concept, reachability assessment, exploitability reasoning, false-positive assessment, mitigation recommendation, patch verification, and regression test results.

Recommendation 5: Add risk indicators and near-miss reporting

Providers should define indicators for emerging or materialized cybersecurity risks posed by AI systems towards AI systems only if it disrupts the proper functionality or operation of the AI system, which might trigger partial or total failure of the AI system at hand. These indicators should include, where relevant, abnormal input patterns, unexplained output changes, model

performance degradation, data distribution shifts, unauthorized model or data modifications, prompt-injection attempts, system-prompt leakage attempts, excessive inference probing, anomalous tool calls, unexpected agent behavior, and integrity verification failures.

Systematic benchmarking work on prompt injection and broader work on ML lifecycle assurance support the need for observable indicators, test evidence, and post-deployment monitoring rather than relying only on pre-deployment assessment [9, 12].

Recommendation 6: Strengthen safety-critical assurance

For AI systems used in safety-critical or critical-infrastructure contexts, providers and deployers should apply additional assurance measures proportionate to the potential impact on safety, security, and operational continuity. Such measures should include independent verification and validation, strict configuration management, formal change control, red-team testing, incident escalation, human-in-command oversight, fail-safe design, and integration with sector-specific safety and cybersecurity requirements.
Research on ML lifecycle assurance and ML safety similarly emphasises the need for lifecycle evidence, traceability, monitoring, and safety-oriented controls when machine-learning systems are deployed in consequential contexts [12, 13].

Conclusion

The draft standard should be supported as an important foundation for the cybersecurity of AI systems covered under the EU AI Act. Its lifecycle structure, AI-specific threat taxonomy, testing provisions, and documentation requirements are valuable and broadly aligned with responsible computing principles. It provides a useful technical basis for identifying vulnerabilities and threats affecting AI systems, selecting proportionate cybersecurity measures, and documenting the evidence needed for assurance and review.

In this response, “AI-specific cybersecurity” refers to vulnerabilities and threat vectors that can compromise an AI system because of its AI-related components, functions, and data flows. These include manipulation of training, fine-tuning, validation, inference, or retrieval data; compromise of models, configurations, prompts, embedded instructions, external tools, interfaces, or agent permissions; adversarial inputs; model extraction; and unauthorised modification of system behaviour [1, 2, 4, 5, 7, 8]. These concerns extend conventional cybersecurity risks because compromise of data, models, connectors, permissions, or deployment configurations can alter an AI system’s behaviour, outputs, or actions through its learning, inference, retrieval, or agentic capabilities.

This is distinct from the possible consequences of a cybersecurity incident involving an AI system. Depending on the intended purpose and deployment context, compromised AI outputs or actions may affect organisations, safety, security, operational continuity, the environment,

essential services, or other legally protected interests. Such consequences should inform risk determination, testing priorities, incident response, and residual-risk acceptance, while the standard remains focused on cybersecurity protection of the AI system itself and coherent with related AI risk-management standards such as prEN 18228.

The standard should therefore be strengthened to reflect the rapid evolution of AI systems into operational, agentic, and supply-chain-dependent systems. It should more explicitly address AI supply-chain security; treat data governance as a cybersecurity requirement; add controls for retrieval-augmented, tool-connected, and agentic AI systems; recognise AI-assisted cybersecurity testing as an emerging practice requiring reproducibility, validation, and containment; and define indicators for emerging risks, incidents, and near misses. The growing literature on AI supply chains, prompt injection, tool-integrated agents, poisoning attacks, model confidentiality, and AI-assisted vulnerability detection supports the need for these additions [2, 3, 6, 7, 8, 9, 10, 11].

The recommendations should be understood as practical guidance intended to improve the usability and technical effectiveness of the draft standard, not to expand prEN 18282 into a general social-risk or safety standard. They should help organisations answer a concrete cybersecurity question: whether the AI system, together with its data, models, tools, integrations, externally supplied components, and operational dependencies, is sufficiently protected against evolving cyber threats.

For consequential deployments, the standard should support proportionate additional assurance. Where the intended purpose, deployment context, access privileges, or downstream reliance on AI outputs could affect safety, security, operational continuity, essential services, or other significant interests, the AI system should be able to enter a safe or controlled state, recover from a verified baseline, provide sufficient traceability of what occurred, and ensure that any residual cybersecurity risk is accepted through an accountable decision. This approach is consistent with lifecycle assurance and machine-learning safety literature, which emphasises evidence, traceability, monitoring, and context-sensitive controls across development and deployment [12, 13].

Frontier AI capabilities further reinforce the urgency of this work. AI systems are increasingly able to assist vulnerability discovery, test generation, exploitability analysis, patching, and defensive cybersecurity workflows. These capabilities can strengthen defence when properly governed, but they may also reduce the time and expertise required to identify and exploit weaknesses. Standards should therefore not only ask whether AI systems are cybersecure and resilient, but also whether organisations are prepared to govern AI-enabled cybersecurity, AI supply-chain dependencies, and agentic operational access in a controlled, evidence-based, and proportionate manner.

Main References

[1] Hu, Yupeng, et al. “Artificial Intelligence Security: Threats and Countermeasures.” *ACM Computing Surveys*, vol. 55, no. 1, 2023, article 20. https://doi.org/10.1145/3487890.

[2] Wang, Xuwang, et al. “Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems.” *ACM Computing Surveys*, vol. 55, no. 7, 2023, article 133. https://doi.org/10.1145/3538707.

[3] Wang, Shenao, Yanjie Zhao, Xinyi Hou, and Haoyu Wang. “Large Language Model Supply Chain: A Research Agenda.” *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 5, 2025, article 147. https://doi.org/10.1145/3708531.

[4] Shokri, Reza, et al. “Membership Inference Attacks Against Machine Learning Models.” *Proceedings of the 2017 IEEE Symposium on Security and Privacy*, IEEE, 2017, pp. 3–18. https://doi.org/10.1109/SP.2017.41.

[5] Carlini, Nicholas, et al. “Extracting Training Data from Large Language Models.” *Proceedings of the 30th USENIX Security Symposium*, USENIX Association, 2021, pp. 2633–2650.

[6] Sood, Aditya K., and Sherali Zeadally. “Malicious AI Models Undermine Software Supply-Chain Security.” *Communications of the ACM*, vol. 68, no. 6, 2025, pp. 62–71. https://doi.org/10.1145/3704724.

[7] Abdelnabi, Sahar, et al. “Not What You’ve Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection.” *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, ACM, 2023, pp. 79–90. https://doi.org/10.1145/3605764.3623985.

[8] Zhan, Qiusi, Zhixiang Liang, Zifan Ying, and Daniel Kang. “InjecAgent: Benchmarking Indirect Prompt Injections in Tool-Integrated Large Language Model Agents.” *Findings of the Association for Computational Linguistics: ACL 2024*, Association for Computational Linguistics, 2024, pp. 10471–10506. https://doi.org/10.18653/v1/2024.findings-acl.624.

[9] Liu, Yupei, et al. “Formalizing and Benchmarking Prompt Injection Attacks and Defenses.” *Proceedings of the 33rd USENIX Security Symposium*, USENIX Association, 2024, pp. 1831–1847.

[10] He, Jingxuan, and Martin Vechev. “Large Language Models for Code: Security Hardening and Adversarial Testing.” *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, ACM, 2023, pp. 1865–1879. https://doi.org/10.1145/3576915.3623175.

[11] Zhou, Xin, Sicong Cao, Xiaobing Sun, and David Lo. “Large Language Model for Vulnerability Detection and Repair: Literature Review and the Road Ahead.” *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 5, 2025, article 145. https://doi.org/10.1145/3708522.

[12] Ashmore, Rob, Radu Calinescu, and Colin Paterson. “Assuring the Machine Learning Lifecycle: Desiderata, Methods, and Challenges.” *ACM Computing Surveys*, vol. 54, no. 5, 2021, article 111. https://doi.org/10.1145/3453444.

[13] Mohseni, Sina, et al. “Taxonomy of Machine Learning Safety: A Survey and Primer.” *ACM Computing Surveys*, vol. 55, no. 8, 2023, article 157. https://doi.org/10.1145/3551385.

Additional Current Practice and Policy References Consulted

ACM Europe Technology Policy Committee. “Calibrating Oversight for Agentic Frontier Models.” *Association for Computing Machinery*, 2026, https://www.acm.org/public-policy/europe-tpc/calibrating-oversight-agentic-frontier-models-04272026.

Broadcom. “What We Learned Testing Frontier AI Security Models Against Our Own Code.” *Broadcom*, 2026, https://news.broadcom.com/security/frontier-ai-security-models-code-testing-results.

CEN/CENELEC. *prEN 18282: Artificial Intelligence: Cybersecurity Specifications for AI Systems*. Draft European Standard, 2026.

Cloudflare. “Project Glasswing: What Mythos Showed Us.” *Cloudflare Blog*, 2026, https://blog.cloudflare.com/cyber-frontier-models/.

European Commission. “Commission Presents EU Action Plan on Cybersecurity and Artificial Intelligence.” *European Commission Press Corner*, 7 July 2026, https://ec.europa.eu/commission/presscorner/detail/en/ip_26_1544.

Mozilla. “Behind the Scenes: Hardening Firefox with Claude Mythos Preview.” *Mozilla Hacks*, 2026, https://hacks.mozilla.org/2026/05/behind-the-scenes-hardening-firefox/.

National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework*. NIST, 2023, https://www.nist.gov/itl/ai-risk-management-framework.

OWASP. *OWASP Top 10 for Large Language Model Applications*. OWASP Foundation, 2025, https://owasp.org/www-project-top-10-for-large-language-model-applications/.

Palo Alto Networks. “Defender’s Guide to the Frontier AI Impact on Cybersecurity: May 2026 Update.” *Palo Alto Networks*, 2026, https://www.paloaltonetworks.com/blog/2026/05/defenders-guide-frontier-ai-impact-cybersecurity-may-2026-update/.

Stenberg, Daniel. “Mythos Finds a Curl Vulnerability.” *Daniel.haxx.se*, 11 May 2026, https://daniel.haxx.se/blog/2026/05/11/mythos-finds-a-curl-vulnerability/.

[i] This policy article is based on the recommendations shared earlier with the Belgian Mirror in response to draft prEN 18282:2026 Cybersecurity Specifications for AI Systems.

[ii] Authors: Ahmed Nagy, Chris Hankin; Reviewers: Francisco Medeiros; Editor: Tom Romanoff