International Standard

ISO/IEC 27090

Cybersecurity — Artificial intelligence (AI) — Addressing security threats and compromises to AI systems

Cybersécurité — Intelligence artificielle (IA) — Traitement des menaces et des situations susceptibles de compromettre les systèmes d'IA

First edition
2026-10

Reference number
ISO/IEC 27090:2026(en)

© ISO/IEC 2026

Contents

Page

Foreword ..... **v**
Introduction ..... **vi**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Abbreviated terms **5**
5 Application of information security **5**
5.1 General 5
5.2 AI governance programme 6
5.3 Zero trust principles 7
5.4 Applying software engineering best practices to AI 8
5.5 Supply chain management 8
5.5.1 General 8
5.5.2 AI-specific supply chain threats 9
5.5.3 Complex AI supply chains 9
5.5.4 Specific attack vectors 9
5.5.5 Mitigation strategies 9
5.6 Threat modelling 10
5.7 Red teaming 10
6 Types of potential compromise to AI systems **10**
6.1 General 10
6.2 Data poisoning attack 13
6.2.1 Description 13
6.2.2 Failures resulting from data poisoning attacks 14
6.2.3 Detection of data poisoning attacks 14
6.2.4 Mitigation of data poisoning attacks 15
6.3 Evasion attack 17
6.3.1 Description 17
6.3.2 Failures resulting from evasion attacks 18
6.3.3 Detection of evasion attacks 18
6.3.4 Mitigation of evasion attacks 18
6.4 Membership inference 20
6.4.1 Description 20
6.4.2 Failures resulting from membership inference 20
6.4.3 Detection of membership inference 21
6.4.4 Mitigation of membership inference 21
6.5 Model exfiltration 21
6.5.1 Description 21
6.5.2 Failures resulting from model exfiltration 22
6.5.3 Detection of model exfiltration 22
6.5.4 Mitigation of model exfiltration 23
6.6 Model inversion 23
6.6.1 Description 23
6.6.2 Failures resulting from model inversion 23
6.6.3 Detection of model inversion 24
6.6.4 Mitigation of model inversion 24
6.7 Direct model poisoning 24
6.7.1 Description 24
6.7.2 Failures resulting from direct model poisoning 25
6.7.3 Detection of direct model poisoning 25
6.7.4 Mitigation of direct model poisoning 25
6.8 Sensitive model output 25
6.8.1 Description 25

© ISO/IEC 2026 – All rights reserved

6.8.2 Failures resulting from sensitive model output ... 25
6.8.3 Detection of sensitive model output ... 25
6.8.4 Mitigations of sensitive model output ... 25
6.9 Prompt injection ... 26
6.9.1 Description ... 26
6.9.2 Failures resulting from prompt injection ... 26
6.9.3 Detection of prompt injection ... 26
6.9.4 Mitigation of prompt injection ... 27
6.10 Output contains injection attacks ... 27
6.10.1 Description ... 27
6.10.2 Failures resulting from output contains injection attacks ... 27
6.10.3 Detection of output contains injection attacks ... 27
6.10.4 Mitigations of output contains injection attacks ... 27
6.11 Considerations for AI systems ... 28
6.11.1 General ... 28
6.11.2 Reduce impact by data limitation ... 28
6.11.3 Governing and controlling the effect of unwanted model behaviour during the operation stage ... 29
6.11.4 AI system and data provenance ... 32

7 Mitigations and their interactions with threats and other applied mitigations ... 33
7.1 General ... 33
7.2 Conflicting interactions of mitigations ... 33
7.3 Continuity of mitigations across the AI life cycle ... 33
7.4 Model and mitigation deterioration over time ... 33
7.5 Logging and monitoring ... 34
7.6 Development environment protection ... 34
7.7 Malicious input detection and filtering ... 35
7.7.1 Detecting out-of-distribution inputs ... 35
7.7.2 Detecting adversarial inputs ... 35
7.8 Throttling model use ... 36
7.9 Optimising the size of the model ... 36

Annex A (informative) **Mapping attacks to AI system life cycle and to assets ... 37**

Annex B (informative) **AI-specific versions of conventional attacks ... 44**

Annex C (informative) **Federated machine learning ... 49**

Bibliography ... 51

© ISO/IEC 2026 – All rights reserved

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

ISO and IEC draw attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO and IEC take no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO and IEC had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents and https://patents.iec.ch. ISO and IEC shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Joint Technical Committee ISO/IEC JTC 1, *Information technology*, Subcommittee SC 27, *Information security, cybersecurity and privacy protection*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

© ISO/IEC 2026 – All rights reserved

Introduction

Mitigating the risks of security threats to AI systems in a timely manner can help to improve the trustworthiness of AI systems and of the organizations that use or develop them, and thus earn trust in these systems and organizations. As technology continues to advance and innovate at a fast pace, emerging and new security vulnerabilities discovered in AI systems can require continuous evaluation of security objectives.

AI systems are subject to all of the traditional cybersecurity vulnerabilities, exploits and attacks. They can be subject to new threats, notably those based on data, input manipulation or actions of AI models. However, these additional concerns do not supersede conventional cybersecurity concerns. As the overall cybersecurity threat landscape evolves, the need to protect AI systems against non-AI system specific threats continues to be an ongoing pursuit.

AI systems have become more widespread as organizations embrace digital transformation, resulting in an increasing likelihood of security attacks on these systems. Furthermore, as a result of the disparate and widespread use of AI systems, and especially within safety critical contexts, the result of security attacks can be severe and in some cases, lead to physical or mental repercussions for individuals.

Due to the way AI systems are developed, and their strong dependency on training data, which can include user data or unvetted data, these systems can have additional vulnerabilities compared to conventional information processing systems. Furthermore, there is a lack of understanding when it comes to the inner workings of AI, specifically when it comes to understanding and diagnosing individual instances of AI behaviour. This document provides guidance to organizations who are choosing to develop or deploy AI systems, providing awareness of the types of security threats such systems can face, and helping organizations understand the consequences of such threats and the most commonly related detection and mitigation techniques. Such awareness and understanding of this new threat landscape is imperative for securing AI systems against adversarial attacks.

This document describes various security threats and the exploitation of vulnerabilities that can happen when using AI systems. The objective is to identify recommended measures to improve the security of AI systems by identifying threats and the consequences of these threats. The identification of threats can help users and developers of AI systems better understand security issues related to AI systems.

Security threats and vulnerabilities with a lack of controls in AI systems can lead to failures, impacting both individuals and businesses.

It can be challenging to detect some threats to AI systems and thus, developing controls without effective threat detection would not be useful or adequate. This document can help identify where security threats can be effectively addressed.

In general, it is recommended that the established commonly accepted security practices in ISO/IEC 27001[1] and ISO/IEC 27002[2] are adhered to. However, the reliance of many AI systems on training data introduces new threats. Attacks such as data poisoning can cause undesirable outcomes by injecting malicious or unwanted data into training data. During the operational phase, AI systems gather data to support future model updates. However, this process can be vulnerable to the introduction of corrupted data, potentially compromising the integrity of future model retraining during the design and development stage. Where continuous or iterative learning occurs, data inputs supplied during operation can also become training data, so corrupted inputs can also become a vector for data poisoning. Accordingly, AI or machine-learning-enabled work, data, and control flows can benefit from some additional controls. This document can provide the necessary insights to apply both established and new controls, for example by explaining what the assets are, and what aspects need protecting.

Other attacks such as model exfiltration through querying can be mitigated by conventional information security measures such as rate limiting. Models and intellectual property can be stolen even when effective conventional information security measures are in place, since a model exfiltration attack is established through legitimate querying of the model. Threat detection is a challenge but an essential component to effectively mitigate model exfiltration. The measures used to mitigate AI threats can use a defence-in-depth strategy that leverages multiple security measures to protect an organization's assets.

© ISO/IEC 2026 – All rights reserved

If a choice is made to use unvetted or unverified data for training, risks related to data poisoning or attacks that can exploit ML data or control flows can increase.

© ISO/IEC 2026 – All rights reserved

Cybersecurity — Artificial intelligence (AI) — Addressing security threats and compromises to AI systems

1 Scope

This document addresses security threats and compromises specific to artificial intelligence (AI) systems. This document aims to provide information to organizations to help them better understand the consequences of security threats specific to AI systems, throughout their life cycle, and descriptions of how to detect and mitigate such threats.

This document is applicable to all types and sizes of organizations, including public and private companies, government entities and not-for-profit organizations that develop or use AI systems.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO/IEC 22989, *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)

Bibliography

[1] ISO/IEC 27001:2022, *Information security, cybersecurity and privacy protection — Information security management systems — Requirements*

[2] ISO/IEC 27002:2022, *Information security, cybersecurity and privacy protection — Information security controls*

[3] ISO/IEC 27000:2026, *Information technology — Security techniques — Information security management systems — Overview and vocabulary*

[4] ISO/IEC 27033-1:2015, *Information technology — Security techniques — Network security — Part 1: Overview and concepts*

[5] ISO/IEC 25030:2007[2], *Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Quality requirements*

[6] ISO 22300:2025, *Security and resilience — Vocabulary*

[7] NIST SP 800-12, *An Introduction to Information Security, June 2017*[3]

[8] ISO/IEC 5259-1:2024, *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 1: Overview, terminology, and examples*

[9] ISO/PAS 8800:2024, *Road vehicles — Safety and artificial intelligence*

[10] ISO/IEC 27071:2023, *Cybersecurity — Security recommendations for establishing trusted connections between devices and services*

[11] ISO/IEC 5338:2023, *Information technology — Artificial intelligence — AI system life cycle processes*

[12] ISO/IEC 23894:2023, *Information technology — Artificial intelligence — Guidance on risk management*

[13] ETSI GR SAI 001 V1.1.1, *Securing Artificial Intelligence (SAI); AI Threat Ontology*. [viewed on 2023-07-19]. Available at: https://www.etsi.org/deliver/etsi_gr/SAI/001_099/001/01.01.01_60/gr_SAI001v010101p.pdf

[14] ETSI EN 304-223, *V2.1.1, Securing Artificial Intelligence (SAI), Baseline Cyber Security Requirements for AI Models and Systems.*

[15] ETSI TR 104 128, *V1.1.1, Securing Artificial Intelligence (SAI); Guide to Cyber Security for AI Models and Systems.*

[16] ISO/IEC 42001:2023, *Information technology — Artificial intelligence — Management system*

[17] NIST SP 800-218, *Secure Software Development Framework (SSDF) Version 1.1: Recommendations for Mitigating the Risk of Software Vulnerabilities, February 2022*[4]

[18] National Cyber Security Centre UK. *Guidelines for secure AI system development*, 27 November 2023. [viewed on 2026-02-18]. Available at https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

[19] ISO/IEC 38507:2022, *Information technology — Governance of IT — Governance implications of the use of artificial intelligence by organizations*

[20] ISO/IEC/IEEE 12207, *Systems and software engineering — Software life cycle processes*

2) Cancelled and replaced by ISO/IEC 25030:2019.

3) Available at: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-12r1.pdf

4) Available at https://nvlpubs.nist.gov/nistpubs/specialpublications/nist.sp.800-218.pdf

© ISO/IEC 2026 – All rights reserved

[21] ISO/IEC 5181, *Information technology — Security and privacy — Data provenance*

[22] ISO/IEC TS 42119-2:2025, *Artificial intelligence — Testing of AI — Part 2: Overview of testing AI systems*

[23] ISO/IEC TS 42119-3, *Artificial intelligence — Testing of AI — Part 3: Verification and validation analysis of AI systems*

[24] *Threat Modeling AI/ML Systems and Dependencies*. Microsoft. [viewed on 2026-02-18]. Available at: https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml

[25] Conklin L. et al. *OWASP Threat modelling process*. [viewed on 2026-02-18]. Available at https://owasp.org/www-community/Threat_Modeling_Process

[26] Nelson B. et al. *Exploiting Machine Learning to Subvert Your Spam Filter*. [viewed on 2026-02-18]. Available at https://www.usenix.org/legacy/event/leet08/tech/full_papers/nelson/nelson.pdf

[27] Xiao H. et al. *Is Feature Selection Secure against Training Data Poisoning?* 2018. Available from https://arxiv.org/abs/1804.07933

[28] Andrew Newell et al. *On the Practicality of Integrity Attacks on Document-Level Sentiment Analysis*. 2014. Available at https://nds2.ccs.neu.edu/papers/aml_aisec2014.pdf

[29] Roei Schuster et al. *You Autocomplete Me: Poisoning Vulnerabilities in Neural Code Completion*. 2020. Available at https://arxiv.org/pdf/2007.02220.pdf

[30] Zhang H. et al. *Practical Data Poisoning Attack against Next-Item Recommendation*. 2020. Available at https://arxiv.org/pdf/2004.03728.pdf

[31] *MITRE ATLAS Case Studies.* [viewed on 2026-02-18]. https://atlas.mitre.org/studies

[32] Bursztein E. *Attacks against machine learning — an overview*. May 2018. [viewed on 2026-02-18]. Available at https://elie.net/blog/ai/attacks-against-machine-learning-an-overview/

[33] ETSI, *Specialist Task Force 584: Artificial Intelligence for IoT Systems.*https://portal.etsi.org/STF/STFs/STF-Homepages/STF584

[34] Narla, A., Kuprel, B., Sarin, K. Y., Novoa, R. A., & Ko, J. (2018). *Automated Classification of skin lesions: From pixels to practice*. Journal of Investigative Dermatology, 138(10), 2108–2110. https://doi.org/10.1016/j.jid.2018.06.175

[35] ISO/IEC 8183:2023, *Information technology — Artificial intelligence — Data life cycle framework*

[36] ISO/IEC/TS 6254, *Information technology — Artificial intelligence — Objectives and approaches for explainability and interpretability of machine learning (ML) models and artificial intelligence (AI) systems*

[37] Chen B., Carvalho W., Baracaldo N., Ludwig H., Edwards B., Lee T. et al. (2018). *Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering*. CoRR, abs/1811.03728. Available at https://arxiv.org/abs/1811.03728

[38] ISO 22316, *Security and resilience — Organizational resilience — Principles and attributes*

[39] Diakonikolas I., Kamath G., Kane D., Li J., Steinhardt J., Stewart A. Sever. *A robust meta-algorithm for stochastic optimization.* In International Conference on Machine Learning, pages 1596–1606. PMLR, 2019. Available at: https://arxiv.org/abs/1803.02815, doi:10.48550/arXiv.1803.02815.

[40] Alexander Levine and Soheil Feizi. *Deep partition aggregation: Provable defenses against general poisoning attacks*. In International Conference on Learning Representations, 2021. URL: https://openreview.net/forum?id=YUGG2tFuPM, doi:10.48550/arXiv.2006.14768

[41] Shumailov, Ilia, et al. *Manipulating sgd with data ordering attacks*. Advances in Neural Information Processing Systems. 34 (2021): 18021-18032

[42] ISO/IEC 5259 (all parts), *Artificial intelligence — Data quality for analytics and machine learning (ML)*

© ISO/IEC 2026 – All rights reserved

[43] Liu K. Brendan Dolan-Gavitt, Siddharth Garg. *Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks*, Available at https://arxiv.org/pdf/1805.12185.pdf

[44] Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy. *Explaining and Harnessing Adversarial Examples*. ICLR conference, 2015

[45] Chen S.-T., Cornelius C., Martin J., Horng D. (Polo) Chau. *ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector*, ECML/PKDD (1) 2018: 52-68, (https://link.springer.com/chapter/10.1007/978-3-030-10997-4_50).

[46] Besnier V., Bursuc A., Picard D., Briot A. *Triggering Failures: Out-Of-Distribution detection by learning from local adversarial attacks in Semantic Segmentation.* arXiv, 2021. Available from https://arxiv.org/abs/2108.01634

[47] Sharif M., Bhagavatula S., Bauer L., Reiter M.K. *Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition*. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016

[48] Biggio, B. and Roli, F. *Wild patterns: Ten years after the rise of adversarial machine learning*, *Pattern Recognition*, 2018

[49] Papernot, N. *Characterizing the Limits and Defenses of Machine Learning in Adversarial Settings*, PennState University Libraries, 2018

[50] Huang S., Papernot N., Goodfellow I., Duan Y., Abbeel P. (2017) *Adversarial attacks on neural network policies.* arXiv preprint arXiv: 1702.02284. Available at:https://arxiv.org/abs/1702.02284

[51] Zhanhao H., Siyuan H., Xiaopei Z., Fuchun S., Bo Z., Xiaolin H. *Adversarial Texture for Fooling Person Detectors in the Physical World*, arXiv, 2022. Available at https://arxiv.org/pdf/2203.03373

[52] Meng, S., Zelin, L., Liehuang, Z., Ke, X. and Xiaojiang, D. VLA: A *Practical Visible Light-Based Attack on Face Recognition Systems in Physical World*, Proceedings ACM Interact. Mob. Wearable Ubiquitous Technol., 2019

[53] Gnanasambandam A., Sherman A.M., Chan S.H. *Optical Adversarial Attack*, arXiv, 2021. Available at https://arxiv.org/pdf/2108.06247

[54] Tencent Keen Security Lab. *Experimental Security Research of Tesla Autopilot*. 2019. [viewed on 2026-02-18] https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf

[55] Biggio, B., Corona, I., He, Z.M., Chan P., Giacinto, G., Yeung, D.S. and Roli,F. *One-and-a-Half-Class Multiple Classifier Systems for Secure Learning Against Evasion Attacks at Test Time*, MCS, 2015

[56] Meng, D. and Chen, H. MagNet: *a Two-Pronged Defense against Adversarial Examples*, arXiv, 2017

[57] Damer N., Moseguí Saladié A., Braun A., Kuijper A. MorGAN: *Recognition Vulnerability and Attack Detectability of Face Morphing Attacks Created by Generative Adversarial Network*, Proceedings of 2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS), 2018

[58] Xu W. David Evans, Yanjun Qi, *Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks*. NDSS 2018 (https://www.ndss-symposium.org/wp-content/uploads/2018/02/ndss2018_03A-4_Xu_paper.pdf).

[59] ISO/IEC TR 24029-1, *Artificial Intelligence (AI) — Assessment of the robustness of neural networks — Part 1: Overview*

[60] ISO/IEC 24029-2, *Artificial intelligence (AI) — Assessment of the robustness of neural networks — Part 2: Methodology for the use of formal methods*

[61] Wu X. et al. *Reinforcing adversarial robustness using model confidence induced by adversarial training*. International Conference on Machine Learning. PMLR, 2018.

© ISO/IEC 2026 – All rights reserved

[62] Bai et al. *Recent Advances in Adversarial Training for Adversarial Robustness.* Available at https://arxiv.org/pdf/2102.01356.pdf

[63] Jeremy M. Cohen, Elan Rosenfeld, J. Zico Kolter, *Certified Adversarial Robustness via Randomized Smoothing*. June 2019. Available at https://arxiv.org/abs/1902.02918

[64] Salman H., Sun M., Yang G., Kapoor A.J. Zico Kolter, *Denoised Smoothing: A Provable Defense for Pretrained Classifiers*, September 2020. Available at https://arxiv.org/abs/2003.01908

[65] Gao Y. Ilia Shumailov, Kassem Fawaz, Nicolas Papernot. *On the Limitations of Stochastic Pre-processing Defenses.* Proceedings of the 36th Conference on Neural Information Processing Systems. October 2022. Available at https://arxiv.org/abs/2206.09491

[66] Liao F., Liang M., Dong Y., Pang T., Hu X., Zhu J. *Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser*, May 2018. Available at https://arxiv.org/abs/1712.02976

[67] Maho T., Furon T., Le Merrer E. *Randomized Smoothing under Attack: How Good is it in Practice?* April 2022. Available at https://arxiv.org/abs/2204.14187

[68] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, Kaiming He, *Feature Denoising for Improving Adversarial Robustness*. Available at https://doi.org/10.48550/arXiv.1812.03411

[69] ENISA. *Securing machine learning algorithms*. [viewed on 2026-02-18]. Available at: https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

[70] Papernot, Nicolas, et al. *Distillation as a defense to adversarial perturbations against deep neural networks*. 2016 IEEE symposium on security and privacy (SP). IEEE, 2016.

[71] Shokri R., Stronati M., Song C., Shmatikov V. *Membership inference attacks against machine learning models*, 2017. Available at https://arxiv.org/pdf/1610.05820

[72] Salem A., Zhang Y., Humbert M., Berrang P., Fritz M., Backes M. *ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models*. 2018. Available at: https://arxiv.org/abs/1806.01246

[73] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, Somesh Jha. *Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting*. Available at https://arxiv.org/pdf/1709.01604

[74] Ye J., Maddi A., Murakonda S.K., Bindschaedler V., Shokri R. *Enhanced membership inference attacks against machine learning models.* In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pages 3093–3106, 2022

[75] Ateniese G., Felici G., Mancini L.V., Spognardi A., Villani A., Vitali D. *Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers*, 2013. arXiv:1306.4447. Available at: https://arxiv.org/abs/1306.4447

[76] ISO/IEC 27701:2025, *Information security, cybersecurity and privacy protection — Privacy information management systems — Requirements and guidance*

[77] Shokri R. et al. *Membership Inference Attacks Against Machine Learning Models*. 2017. Available at https://arxiv.org/pdf/1610.05820.pdf

[78] Carlini N. et al. Extracting Training Data from Diffusion Models. 2023. Available at https://arxiv.org/pdf/2301.13188.pdf

[79] Abadi M. et al. *Deep learning with differential privacy*. Proc. ACM Conf. Computer and Communications Security, 2016

[80] Jayaraman B., Evans D. *Evaluating differentially private machine learning in practice*. Proc. USENIX Security, 2019

[81] Mitigation Strategy Report. *Securing Artificial Intelligence (SAI)*; ETSI GR SAI 005 V1.1.1. 2021-03.

© ISO/IEC 2026 – All rights reserved

[82] Kesarwani M., Mukhoty B., Arya V., Mehta S. (2017, November 20). *Model extraction warning in MLAAS paradigm*. arXiv.org. https://arxiv.org/abs/1711.07221

[83] Pal S., Gupta Y., Kanade A., Shevade S. (2021, July 12). *Stateful detection of model extraction attacks*. arXiv.org. https://arxiv.org/abs/2107.05166

[84] Tramèr F., Zhang F., Juels A., Reiter M.K., Ristenpart T. *Stealing Machine Learning Models via Prediction APIs*. 25th USENIX Security Symposium, 2016

[85] Juuti M., Szyller S., Marchal S., Asokan N. *PRADA: Protecting Against DNN Model Stealing Attacks.* 2019 IEEE European Symposium on Security and Privacy (EuroS&P), Stockholm, Sweden, 2019, pp. 512-527, doi: 10.1109/EuroSP.2019.00044. available at https://ieeexplore.ieee.org/document/8806737

[86] Atli, B.G., Szyller, S., Juuti, M., Marchal, S., Asokan, N. (2020). *Extraction of Complex DNN Models: Real Threat or Boogeyman?.* In: Shehory, O., Farchi, E., Barash, G. (eds) Engineering Dependable and Secure Machine Learning Systems. EDSMLS 2020. Communications in Computer and Information Science, vol 1272. Springer, Cham. https://doi.org/10.1007/978-3-030-62144-5_4

[87] Fredrikson M., Jha S., Ristenpart T. *Model inversion attacks that exploit confidence information and basic countermeasures*, 2015. https://dl.acm.org/doi/pdf/10.1145/2810103.2813677

[88] McSherry F. *Statistical inference considered harmful*. 2016, Github. [viewed on 2026-02-18] https://github.com/frankmcsherry/blog/blob/master/posts/2016-06-14.md

[89] Veale M., Binns R., Edwards L. Algorithms that remember: model inversion attacks and data protection law. October 2018. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences Volume 376, Issue 213. https://doi.org/10.1098/rsta.2018.0083

[90] Carlini N. et al. *Extracting Training Data from Large Language Models*. 2012. Available at https://arxiv.org/abs/2012.07805

[91] Lei Y. (Murray) and Miao, Sentao and Momot, Ruslan, *Privacy-Preserving Personalized Revenue Management.* October 2020. Available at SSRN: https://ssrn.com/abstract=3704446 or http://dx.doi.org/10.2139/ssrn.3704446

[92] ISO/IEC 20889:2018, *Privacy enhancing data de-identification terminology and classification of techniques*

[93] *Adversarial Threat Landscape for Artificial-Intelligence Systems*. MITRE ATLAS. [viewed on 2026-02-18]. Available at: https://atlas.mitre.org/

[94] Mansi Phute, Alec Helbling, Matthew Hull, Shengyun Peng, Sebastian Szyller, Cory Cornelius, Duen Horng Chau. *LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked*. Tiny Papers @ ICLR 2024 (https://arxiv.org/abs/2308.07308).

[95] ISO/IEC 27091, *Cybersecurity and Privacy — Artificial Intelligence — Privacy protection*

[96] ISO/IEC 42105, *Information technology — Artificial intelligence — Guidance for human oversight of AI systems*

[97] ISO/IEC 12792:2025, *Information technology — Artificial intelligence (AI) — Transparency taxonomy of AI systems*

[98] ISO/IEC TS 12791, *Information technology — Artificial intelligence — Treatment of unwanted bias in classification and regression machine learning tasks*

[99] ISO/IEC 21617-1, *Information technology — JPEG Trust — Part 1: Core foundation*

[100] *C2PA Coalition for Content Provenance and Authenticity. C2PA Specification 1.3 – AI/ML Guidance*. C2PA, 2024. Available at:https://c2pa.org/specifications/specifications/1.3/ai-ml/ai_ml.html

© ISO/IEC 2026 – All rights reserved

[101] Raji I.D., Hodas N., Shih C. et al. *Model Provenance for Generative AI.* arXiv preprint arXiv: 2310.16787, October 2023. Available at:https://arxiv.org/pdf/2310.16787

[102] *NIST NCCoE Dioptra* Post Workshop Conference Webinar Recording, Timecode: 30:36, https://cdnapisec.kaltura.com/index.php/extwidget/preview/partner_id/684682/uiconf_id/31013851/entry_id/1_mssgh30m/embed/dynamic

[103] Szyller, S. and Asokan, N., 2023, June. *Conflicting Interactions Among Protection Mechanisms for Machine Learning Models*. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 12, pp. 15179-15187).

[104] Clarysse J., Hörrmann J., Yang F. 2022. *Why adversarial training can hurt robust accuracy*. arXiv preprint arXiv: 2203.02006. Available athttps://arxiv.org/abs/2203.02006

[105] Vela, D., Sharp, A., Zhang, R. *et al. Temporal quality degradation in AI models*. Sci Rep **12**, 11654 (2022). https://doi.org/10.1038/s41598-022-15245-z

[106] Spoczynski M., Melara M., Szyller S. *Atlas: A Framework for ML Lifecycle Provenance & Transparency*, February 2025, https://arxiv.org/abs/2502.19567

[107] Goodfellow-et-al-2016, Ian Goodfellow, Yoshua Bengio, Aaron Courville. *Deep Learning.* MIT Press 2016. Available at: http://www.deeplearningbook.org

[108] *IEEE Guide for Architectural Framework and Application of Federated Machine Learning*, in IEEE Std 3652.1-2020, vol., no., pp.1-69, 19 March 2021, doi: 10.1109/IEEESTD.2021.9382202.

[109] Confidential computing consortium, *Common Terminology for Confidential Computing*, December 2022. Available at https://confidentialcomputing.io/wp-content/uploads/sites/10/2023/03/Common-Terminology-for-Confidential-Computing.pdf

[110] ISO/IEC 18033-6:2019, *IT Security techniques — Encryption algorithms — Part 6: Homomorphic encryption*

[111] I. Shumailov, Y. Zhao, D. Bates, N. Papernot, R.D. Mullins, and R. Anderson. *Sponge examples: Energy-latency attacks on neural networks*. IEEE EuroS&P'21, pages 212–231, 2021. Available at https://doi.org/10.1109/EuroSP51992.2021.00024.

[112] Dahiya P. Shumailov.I., and R. Anderson. *Machine Learning needs Better Randomness Standards: Randomised Smoothing and PRNG-based attacks*. Available at https://arxiv.org/pdf/2306.14043.pdf, 2023.

[113] ISO/IEC 18031, *Information technology — Security techniques — Random bit generation*

[114] Bober-Irizar M., Shumailov I., Zhao Y., Mullins R., Papernot N. *Architectural Backdoors in Neural Networks*. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 24595–24604, 2023. DOI 10.1109/CVPR52729.2023.02356

[115] Peter Kairouz et. al, *Advances and Open Problems in Federated Learning*, March 2021, (https://arxiv.org/abs/1912.04977).

© ISO/IEC 2026 – All rights reserved