ETSI TR 104 032 V1.1.1 (2024-02)

Securing Artificial Intelligence (SAI);
Traceability of AI Models

Reference
DTR/SAI-004

Keywords
artificial intelligence, cyber security, digital right management, ML watermarking, trustworthy AI

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

This Technical Report (TR) has been produced by ETSI Technical Committee Securing Artificial Intelligence (SAI).

Modal verbs terminology

In the present document "**should**", "**should not**", "**may**", "**need not**", "**will**", "**will not**", "**can**" and "**cannot**" are to be interpreted as described in clause 3.2 of the ETSI Drafting Rules (Verbal forms for the expression of provisions).

"**must**" and "**must not**" are **NOT** allowed in ETSI deliverables except when used in direct citation.

1 Scope

The present document describes the role of traceability in the challenge of Securing AI and explores issues related to sharing and re-using models across tasks and industries. The scope includes threats, and their associated remediations where applicable, to ownership rights of AI creators as well as to verification of models origin. Mitigations can be non-AI-Specific (Digital Right Management applicable to AI) and AI-specific techniques (e.g. ML watermarking) from prevention and detection phases. They can be both model-agnostic and model enhancement techniques. The present document aligns terminology with existing ETSI ISG SAI documents and studies, and references/complements previously studied attacks and remediations (ETSI GR SAI 004 [i.2] and ETSI GR SAI 005 [i.3]). It also gathers industrial and academic feedback on traceability and ownership rights protection and model verification in the context of AI.

2 References

2.1 Normative references

Normative references are not applicable in the present document.

2.2 Informative references

References are either specific (identified by date of publication and/or edition number or version number) or non-specific. For specific references, only the cited version applies. For non-specific references, the latest version of the referenced document (including any amendments) applies.

NOTE: While any hyperlinks included in this clause were valid at the time of publication, ETSI cannot guarantee their long term validity.

The following referenced documents are not necessary for the application of the present document but they assist the user with regard to a particular subject area.

[i.1] OpenAI's GPT-3 Language Model: A Technical Overview.

[i.2] ETSI GR SAI 004: "Securing Artificial Intelligence (SAI); Problem Statement".

[i.3] ETSI GR SAI 005: "Securing Artificial Intelligence (SAI); Mitigation Strategy Report".

[i.4] Artificial Intelligence Market Insights, 2020-2023.

[i.5] BankInfo Security®: "To Combat Rogue AI, Facebook Pitches 'Radioactive Data".

[i.6] McKinney S. M. et al. (2020): "International evaluation of an AI system for breast cancer screening". Nature 577, pp. 89-94.

[i.7] Intellectual Property aspects of Machine Learning, NXP, 2022.

[i.8] Foss-Solbrekk K.: "Three routes to protecting AI systems and their algorithms under IP law: The good, the bad and the ugly". Journal of Intellectual Property Law & Practice 16.3 (2021), pp. 247-258.

[i.9] Hu X., Liang L., Li S., Deng L., Zuo P., Ji Y., Xie X., Ding Y., Liu C., Sherwood T. & Xie Y. 2020: "DeepSniffer: A DNN Model Extraction Framework Based on Learning Architectural Hints". In Proceedings of the Twenty-Fifth International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS '20). Association for Computing Machinery, New York, NY, USA, pp. 385-399.

[i.10] Xu Q., Arafin Md T. & Qu G. (2021): "Security of Neural Networks from Hardware Perspective: A Survey and Beyond". In Proceedings of the 26th Asia and South Pacific Design Automation Conference (ASPDAC '21). Association for Computing Machinery, New York, NY, USA, pp. 449-454.

[i.11] Chakraborty A., Mondal A. & Srivastava A. 2020: "Hardware-assisted intellectual property protection of deep learning models". In Proceedings of the 57th ACM/EDAC/IEEE Design Automation Conference (DAC '20). IEEE™ Press, Article 172, pp. 1-6.

[i.12] Goldstein B.F., Patil V.C., Ferreira V.C., Nery A.S., França F.M.G. & Kundu S.: "Preventing DNN Model IP Theft via Hardware Obfuscation". In IEEE™ Journal on Emerging and Selected Topics in Circuits and Systems, vol. 11, no. 2, pp. 267-277, June 2021.

[i.13] Li J., He Z., Rakin A., Fan D. & Chakrabarti C. (2021): "NeurObfuscator: A Full-stack Obfuscation Tool to Mitigate Neural Architecture Stealing", pp. 248-258. 10.1109/HOST49136.2021.9702279.

[i.14] Mahya Morid A., Alrahis L., Colucci A., Sinanoglu O. & Shafique M. (2022): "NeuroUnlock: Unlocking the Architecture of Obfuscated Deep Neural Networks".

[i.15] Tramer F., Zhang F. J., Reiter M. & Ristenpart T. (2016): "Stealing machine learning models via prediction apis. In 25th USENIX Security Symposium (USENIX Security 16), pp. 601-618.

[i.16] Kálmán S., Al-Afandi, J. & Horváth A. (2019): "MimosaNet: An Unrobust Neural Network Preventing Model Stealing".

[i.17] Uchida Y., Nagai Y., Sakazawa S. & Satoh S. I. (2017): "Embedding watermarks into deep neural networks". In Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, pp. 269-277.

[i.18] Adi Y., Baum C., Cisse M., Pinkas B. & Keshet J. (2018): "Turning your weakness into a strength: Watermarking deep neural networks by backdooring". In 27th USENIX Security Symposium (USENIX Security 18), pp. 1615-1631.

[i.19] Zhang J., Gu Z., Jang J., Wu H., Stoecklin M. P., Huang H. & Molloy I. (2018): "Protecting intellectual property of deep neural networks with watermarking". In Proceedings of the 2018 on Asia Conference on Computer and Communications Security, pp. 159-172.

[i.20] Fan L., Ng K. W. & Chan C. S. (2019): "Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attack. Advances in Neural Information Processing Systems (NIPS)".

[i.21] Chen H., Rouhani B. D., Fu C. Z. & Koushanfar F. (2019): "Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models". In Proceedings of the 2019 on International Conference on Multimedia Retrieval, pp. 105-113.

[i.22] Wang T. and Florian K. (2021) RIGA: "Covert and Robust White-Box Watermarking of Deep Neural Networks". Proceedings of the Web Conference 2021.

[i.23] Li Z., Hu C., Zhang Y. & Guo S. (2019): "How to prove your model belongs to you: a blind-watermark based framework to protect intellectual property of DNN". In Proceedings of the 35th Annual Computer Security Applications Conference, pp. 126-137.

[i.24] Guo J. & Potkonjak M. (2018): "Watermarking deep neural networks for embedded systems". In 2018 IEEE/ACM International Conference on Computer-Aided Design (ICCAD), pp. 1-8.

[i.25] Szyller S., Atli B., Marchal S. & Asokan N. (2019): "DAWN: Dynamic Adversarial Watermarking of Neural Networks".

[i.26] Jia H., Choquette-Choo C. A. & Papernot N. (2020): "Entangled Watermarks as a Defense against Model Extraction".

[i.27] Le Merrer E., Perez P. & Tredan G. (2019): "Adversarial frontier stitching for remote neural network watermarking. Neural Computing and Applications", pp. 1-12.

[i.28] Lukas Nils, Yuxuan Zhang and Florian Kerschbaum. (2019): "Deep neural network fingerprinting by conferrable adversarial examples".

[i.29] Cao Xiaoyu, Jinyuan Jia and Neil Zhenqiang Gong. "IPGuard: Protecting intellectual property of deep neural networks via fingerprinting the classification boundary". Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security.

[i.30] Namba R., & Sakuma J. (2019): "Robust watermarking of neural network with exponential weighting". In Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security, pp. 228-240.

[i.31] Chen Huili, Bita Darvish Rouhani and Farinaz Koushanfar (2019): "BlackMarks: Blackbox Multibit Watermarking for Deep Neural Networks".

[i.32] Sablayrolles A., Douze M., Schmid C. & Jegou H.: "Radioactive data: tracing through training". In Proceedings of the 37th International Conference on Machine Learning, PMLR 119:8326-8335, 2020.

[i.33] Yang Z., Dang H. & Chang E. C. (2019): "Effectiveness of Distillation Attack and Countermeasure on Neural Network Watermarking".

[i.34] Pan S. J. & Qiang Y. (2009): "A survey on transfer learning". IEEE Transactions on knowledge and data engineering, pp. 1345-1359.

[i.35] Hinton G., Oriol V. & Jeff D. (2015): "Distilling the knowledge in a neural network. NIPS Deep Learning and Representation Learning".

[i.36] Papernot N., Song S., Mironov I., Raghunathan A., Talwar K. & Erlingsson Ú. (2018): "Scalable private learning with pate". arXiv:1802.08908.

[i.37] Chen X., Wang W., Bender C., Ding Y., Jia R., Li B. & Song D. (2019): "REFIT: a Unified Watermark Removal Framework for Deep Learning Systems with Limited Data".

[i.38] Chen T., Goodfellow I. & Shlens J. (2016). Net2net: "Accelerating learning via knowledge transfer". In Proceedings of ICLR.

[i.39] Wei T., Wang C., Rui Y. & Chen C. W. (2016): "Network morphism. In International Conference on Machine Learning", pp. 564-572.

[i.40] Wang B., Yao Y., Shan S., Li H., Viswanath B., Zheng H., & Zhao B. (2019): "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks". pp.707-723. 10.1109/SP.2019.00031.

[i.41] Aiken W., Kim H. & Woo S.S. (2020): "Neural Network Laundering: Removing Black-Box Backdoor Watermarks from Deep Neural Networks".

[i.42] Fredrikson M., Jha S. & Ristenpart T. (2015): "Model inversion attacks that exploit confidence information and basic countermeasures". In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322-1333.

[i.43] Kapusta K., Thouvenot V., Bettan O., Beguinet H. & Senet H. (2021): "A Protocol for Secure Verification of Watermarks Embedded into Machine Learning Models". In Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec '21).

[i.44] Kallas K. & Furon T. (2022): "ROSE: A RObust and SEcure DNN Watermarking". 2022 IEEE™ International Workshop on Information Forensics and Security (WIFS), Shanghai, China, 2022, pp. 1-6.

[i.45] Moreau L., Clifford B., Freire J., Futrelle J., Gil Y., Groth P., Kwasnikowska N., Miles S., Missier P., Myers J., Plale B., Simmhan Y., Stephan E. & Van den Bussche J. (2010): "The open provenance model core specification (v1.1)". Future Generation Computer Systems, July 2010.

[i.46] W3C Working Group Note 30 April 2013: "PROV-Overview - An Overview of the PROV Family of Documents".

[i.47] ETSI GR SAI 002: "Securing Artificial Intelligence (SAI); Data Supply Chain Security".

[i.48] Gama J., Žliobaitė I., Bifet A. et al.: "A survey on concept drift adaptation. ACM computing surveys (CSUR)", 2014, 46(4): pp. 1-37.

[i.49] Krawczyk B. & Cano A.: "Online ensemble learning with abstaining classifiers for drifting and noisy data streams. Applied Soft Computing", 2018, 68: pp. 677-692.

[i.50] Kuncheva L. I.: "Change detection in streaming multivariate data using likelihood detectors". IEEE transactions on knowledge and data engineering, 2011, 25(5): pp. 1175-1180.

[i.51] Bifet A. & Gavalda R. (2007): "Learning from time-changing data with adaptive windowing?" Proceedings of the 2007 SIAM international conference on data mining. Society for Industrial and Applied Mathematics, 2007: pp. 443-448.

[i.52] Zhang S., Pan C., Song L. et al. (2021): "Label-Assisted Memory Autoencoder for Unsupervised Out-of-Distribution Detection. Joint European Conference on Machine Learning and Knowledge Discovery in Databases". Springer, Cham, 2021: pp. 795-810.

[i.53] Wu, X., Hu, Z., Pei K, et al. (2021): "Methods for deep learning model failure detection and model adaption: A survey". IEEE™ International Symposium on Software Reliability Engineering Workshops (ISSREW). IEEE™, 2021: 218-223.

[i.54] Mikołajczyk A. & Grochowski M. (2018): "Data augmentation for improving deep learning in image classification problem. International interdisciplinary PhD workshop (IIPhDW)". IEEE, pp. 117-122, 2018.

[i.55] Huang C., Hu Z., Huang X et al. (2021): "Statistical certification of acceptable robustness for neural networks". International Conference on Artificial Neural Networks. Springer, Cham, 2021: pp. 79-90.

[i.56] Shen Z., Cui P., Zhang T. et al. (2020): "Stable learning via sample reweighting". Proceedings of the AAAI Conference on Artificial Intelligence, vol 34(04), pp. 5692-5699.

[i.57] Jo J., Verma V., Bengio Y. (2018): "Modularity matters: Learning invariant relational reasoning tasks".

[i.58] Hummer W., Muthusamy V., Rausch T., Dube P., El Maghraoui K., Murthi A. & Oum P. (2019): "ModelOps: Cloud-Based Lifecycle Management for Reliable and Trusted AI". IEEE™ International Conference on Cloud Engineering (IC2E), pp. 113-120.

[i.59] Vartak M., Subramanyam H., Lee W-E, Viswanathan S., Husnoo S., Madden S. & Zaharia M. (2016): "MODELDB: A System for Machine Learning Model Management. Workshop on Human-In-the-Loop Data Analytics (HILDA)", June 26 2016, San Francisco, CA, USA.

[i.60] Gundersen O. E. (2021): "The fundamental principles of reproducibility. Philosophical Transactions of the Royal Society A", 379(2197), 20200210.

[i.61] Pineau J., Vincent-Lamarre P., Sinha K., Larivière V., Beygelzimer A., d'Alché-Buc F., Fox E. and Larochelle H., 2021: "Improving reproducibility in machine learning research: a report from the NeurIPS 2019 reproducibility program". Journal of Machine Learning Research, 22.

[i.62] Chen B., Wen M., Shi Y., Lin D., Rajbahadur G. K. & Jiang Z. M. (2022, May): "Towards training reproducible deep learning models". In Proceedings of the 44th International Conference on Software Engineering (pp. 2202-2214).

[i.63] Arcuri A. and Briand L.: "A practical guide for using statistical tests to assess randomized algorithms in software engineering". In Proceedings of the 33rd International Conference on Software Engineering, New York, NY, USA, May 2011, pp. 1-10.

[i.64] Pham H.V., Qian S., Wang J., Lutellier T., Rosenthal J., Tan L., Yu Y. & Nagappan, N., 2020, December: "Problems and opportunities in training deep learning software systems: An analysis of variance". In Proceedings of the 35th IEEE/ACM international conference on automated software engineering (pp. 771-783).

[i.65] Elsken T., Metzen J.H. & Hutter F. (2019): "Neural Architecture Search: A Survey".

[i.66] Dehmer M. & Pickl S. (2015): "Network Complexity Measures: an Information-theoretic Approach".

[i.67] M. Fan, W. Wei, X. Xie, Y. Liu, X. Guan and T. Liu. (2021): "Can We Trust Your Explanations? Sanity Checks for Interpreters in Android Malware Analysis". In IEEE™ Transactions on Information Forensics and Security, vol. 16, pp. 838-853.

[i.68] Yan S., Tao G., Liu X. et al. (2020): "Correlations between deep neural network model coverage criteria and model quality". In Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering, New York, NY, USA: Association for Computing Machinery, pp. 775-787.

[i.69] Lyu Y., Rajbahadur G. K., Lin D., Chen B. & Jiang Z. M. (2021): "Towards a consistent interpretation of AIOps models". ACM Transactions on Software Engineering and Methodology (TOSEM), 31(1), 1-38.

[i.70] Tulio Ribeiro M., Singh S. & Guestrin C.: "Why should I trust you?: Explaining the predictions of any classifier". Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM (2016).

[i.71] Tulio Ribeiro M., Singh S. & Guestrin C.:"Anchors: High-Precision Model-Agnostic Explanations" AAAI Conference on Artificial Intelligence (AAAI), 2018.

[i.72] Lundberg, S. M. & Su-In, L. (2017): "A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems".

[i.73] Rajbahadur G. K., Wang S., Oliva G., Kamei Y, Hassan A. E. (2021): "The impact of feature importance methods on the interpretation of defect classifiers". IEEE™ Transactions on Software Engineering, pp. 10.1109/TSE.2021.3056941.

3 Definition of terms, symbols and abbreviations

3.1 Terms

Void.

3.2 Symbols

Void.

3.3 Abbreviations

For the purposes of the present document, the following abbreviations apply:

AI	Artificial Intelligence
AIA	Artificial Intelligence Agent
AIH	Artificial Intelligence Host
API	Application Programming Interface
AUC	Area Under the Curve
DevOps	Development Operations
DNN	Deep Neural Network(s)
DRM	Digital Right Management
FL	Federated Learning
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
IP	Intellectual Property
IPR	Intellectual Property Rights
ML	Machine Learning
MLaaS	Machine Learning as a Service
MLOps	Machine Learning Operations
NDA	Non Disclosure Agreement
OPM	Open Provenance Model
OS	Operating System
PKI	Public Key Infrastructure
TEE	Trusted Execution Environment