ETSI TR 104 048 V1.1.1 (2025-01)

TECHNICAL REPORT

Securing Artificial Intelligence (SAI);
Data Supply Chain Security

Reference

DTR/SAI-007

Keywords

artificial intelligence, security

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 IPR online database.

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™**, **LTE™** and **5G™** logo 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.

Introduction

Artificial Intelligence (AI) and Machine Learning (ML) are fast becoming ubiquitous in almost every sector of society, as AI systems are relied upon to maintain our security, prosperity and health. The compromise of AI systems can therefore have significant impacts on the way of life of vast numbers of people.

However, like any information technology system, AI models are vulnerable to compromise, whether by deliberately hostile or accidental action. One potential vector to compromise AI systems is through the data used to train and operate AI models. If an attacker can introduce incorrect, or incorrectly labelled, data into the model training process, then a model's learning process can be disrupted, and it can be made to produce unintended and potentially harmful results.

This type of attack can be extremely challenging to detect, particularly when, as is increasingly common, the data used to develop and train AI models is part of a complex supply chain. Ensuring the provenance and integrity of the data supply chain will therefore be a key aspect of ensuring the integrity and performance of critical AI-based systems.

The present document has investigated existing mechanisms for carrying out this assurance. AI remains a fast-developing discipline and no legal, policy or standards frameworks have been found that specifically cover data supply chain security. Although many threats can be mitigated by following standard cybersecurity good practice, there is value in producing standards and guidance tailored specifically to AI data supply chains. The conclusion to the present document sets out a number of general principles for consideration in designing and implementing the data supply chain for an AI system.

1 Scope

The present document addresses the security problems arising from data supply chains in in the development of Artificial Intelligence (AI) and Machine Learning (ML) systems. Data is a critical component in the development of AIML systems. Compromising the integrity of data has been demonstrated to be a viable attack vector against such systems (see clause 4). The present document summarizes the methods currently used to source data for training AI, along with a review of existing initiatives for developing data sharing protocols. It then provides a gap analysis on these methods and initiatives to scope possible requirements for standards for ensuring integrity and confidentiality of the shared data, information and feedback.

The present document relates primarily to the security of *data*, rather than the security of models themselves. It is recognized, however, that AI supply chains can be complex and that models can themselves be part of the supply chain, generating new data for onward training purposes. Model security is therefore influenced by, and in turn influences, the security of the data supply chain. Mitigation and detection methods can be similar for data and models, with poisoning of one being detected by analysis of the other.

The present document focuses on security; however, data integrity is not only a security issue. Techniques for assessing and understanding data quality for performance, transparency or ethics purposes are applicable to security assurance too. An adversary aim can be to disrupt or degrade the functionality of a model to achieve a destructive effect. The adoption of mitigations for security purposes will likely improve performance and transparency, and vice versa.

The present document does not discuss data theft, which can be considered a traditional cybersecurity problem. The focus is instead specifically on data manipulation in, and its effect on, AI/ML systems.

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] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, Bo Li: "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning", 2018.

[i.2] Panagiota Kiourti, Kacper Wardega, Susmit Jha, Wenchao Li: "TrojDRL: Evaluation of Backdoor Attacks on Deep Reinforcement Learning", 2020.

[i.3] Kwang-Sung Jun, Lihong Li, Yuzhe Ma, Xiaojin Zhu: "Adversarial Attacks on Stochastic Bandits", 2018.

[i.4] Roei Schuster, Tal Schuster, Yoav Meri, Vitaly Shmatikov: "Humpty Dumpty: Controlling Word Meanings via Corpus Poisoning", 2020.

[i.5] Hengtong Zhang, Tianhang Zheng, Jing Gao, Chenglin Miao, Lu Su, Yaliang Li, Kui Ren: "Data Poisoning Attack against Knowledge Graph Embedding".

[i.6] Mingjie Sun, Jian Tang, Huichen Li, Bo Li, Chaowei Xiao, Yao Chen, Dawn Song: "Data Poisoning Attack against Unsupervised Node Embedding Methods", 2018.

[i.7] Qiang Yang, Yang Liu, Tianjian Chen, Yongxin Tong: "Federated Machine Learning: Concept and Applications", ACM Transactions on Intelligent Systems and Technology, 2019.

[i.8] Arjun Nitin Bhagoji, Supriyo Chakraborty, Seraphin Calo, Prateek Mittal: "Model Poisoning Attacks in Federated Learning. Workshop on Security in Machine Learning at Neural Information Processing Systems", 2018.

[i.9] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, Julien Stainer: "Machine Learning with Adversaries: Byzantine Tolerant Gradient Descent", Advances in Neural Information Processing Systems, 2017.

[i.10] Dong Yin, Yudong Chen, Kannan Ramchandran, Peter Bartlett: "Byzantine-Robust Distributed Learning: Towards Optimal Statistical Rates", International Conference on Machine Learning, 2018.

[i.11] Northrop Grumman, AI Data Supply Chains, 2020.

NOTE: Reference not publicly available.

[i.12] High-Level Expert Group on AI: "Ethics Guidelines for Trustworthy AI", 2019.

[i.13] ETSI TR 104 221: "Securing Artificial Intelligence (SAI); Problem Statement".

NOTE: The above document updates the previously published ETSI GR SAI 004.

[i.14] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, Sharon Xia: "Adversarial Machine Learning - Industry Perspectives", 2020.

[i.15] CESI (China Electronics Standardization Institute): "Artificial Intelligence Standardization White Paper. 2018 edition", 2020, English translation.

[i.16] Microsoft®, MITRE®, et al.: "Adversarial ML Threat Matrix", 2020.

[i.17] Corey Dunn, Nour Mustafa, Benjamin Peter Turnbull: "Robustness Evaluations of Sustainable Machine Learning Models Against Data Poisoning Attacks in the Internet of Things", Sustainability 12(16):6434, 2020.

[i.18] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, Michael Wellman: "SoK: Towards the Science of Security and Privacy in Machine Learning", 2016.

[i.19] Battista Biggio, Fabio Roli: "Wild Patterns, Ten Years After the Rise of Adversarial Machine Learning", 2018.

[i.20] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, Dawn Song: "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning", 2017.

[i.21] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, Debdeep Mukhopadhyay: "Adversarial Attacks and Defenses: A Survey", 2018.

[i.22] Ram Shankar Siva Kumar, Jeffrey Snover, David O'Brien, Kendra Albert, Salome Viljoen: "Failure Modes in Machine Learning", 2019.

[i.23] Andrew Marshall, Jugal Parikh, Emre Kiciman, Ram Shankar Siva Kumar: "Threat Modeling AI/ML Systems and Dependencies", 2019.

[i.24] National Cyber Security Centre: "Supply chain security guidance", 2018.

[i.25] Jon Boyens, Celia Paulsen, Nadya Bartol, Kris Winkler, James Gimbi: "Key Practices in Cyber Supply Chain Risk Management: Observations from Industry". 2021.

[i.26] European Commission: "Joint Press Statement from European Commissioner for Justice Didier Reynders and U.S. Secretary of Commerce Wilbur Ross", 10 August 2020.

[i.27] ETSI TR 104 222: "Securing Artificial Intelligence (SAI); Mitigation Strategy Report".

NOTE: The above document updates the previously published ETSI GR SAI 005.

[i.28] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles Sutton, J.D. Tygar, Kai Xia: "Exploiting Machine Learning to Subvert Your Spam Filter", 2008.

[i.29] Olakunle Ibitoye, Rana Abou-Khamis, Ashraf Matrawy, M. Omair Shafiq: "The Threat of Adversarial Attacks Against Machine Learning in Network Security: A Survey", 2020.

[i.30] Cynthia Rudin: "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead", 2019.

[i.31] ENISA (European Union Agency for Cybersecurity): "Cybersecurity Challenges in the Uptake of Artificial Intelligence in Autonomous Driving", 2021.

[i.32] Bret Cohen, Aaron Lariviere, Tim Tobin: "Understanding the new California Privacy Rights Act: How businesses can comply with the CPRA", 25 November 2020.

[i.33] Ibrahim Hasan: "California Consumer Privacy Act. The Law Society Gazette", 13 July 2020.

[i.34] Linklaters: "Data Protected -- Russia", March 2020.

[i.35] Dora Luo, Yanchen Wang: "China -- Data Protection Overview", OneTrust DataGuidance, November 2020.

[i.36] Tomoki Ishiara: "The Privacy, Data Protection and Cybersecurity Law Review: Japan", October 2020.

[i.37] Linklaters: "Data Protected - Germany", March 2020.

[i.38] Australian Government: "Guide to securing personal information", Office of the Australian Information Commissioner, 5 June 2018.

[i.39] James Walsh: "Security in the supply chain - a post-GDPR approach". Computer Weekly, 7 November 2019.

[i.40] Vyacheslav Khayryuzov: "The Privacy, Data Protection and Cybersecurity Law Review: Russia", 21 October 2020.

[i.41] ETSI TS 119 312: "Electronic Signatures and Infrastructures (ESI); Cryptographic Suites".

[i.42] BSI (Bundesamt für Sicherheit in der Informationstechnik): "Minimum Requirements for Evaluating Side-Channel Attack Resistance of RSA, DSA and Diffie-Hellman Key Exchange Implementations", 2013.

[i.43] Christian Berghoff: "Protecting the integrity of the training procedure of neural networks", 14 May 2020.

[i.44] OpenImages V6.

NOTE: The reference is to a specific version of the OpenImages collection although the collection is regularly updated (https://storage.googleapis.com/openimages/web/download_v6.html).

[i.45] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, Neil Zhenqiang Gong: "Local Model Poisoning Attacks to Byzantine-Robust Federated Learning", 2020.

[i.46] Ilia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A. Erdogdu, Ross Anderson: "Manipulating SGD with Data Ordering Attacks", 2021.

[i.47] Jon-Eric Melsæter.

[i.48] Don DeBold.

[i.49] BSI: "Guidelines for Evaluating Side-Channel and Fault Attack Resistance of Elliptic Curve Implementations".

NOTE: Updated in 2024 from earlier document from 2016 "Minimum Requirements for Evaluating Side-Channel Attack Resistance of Elliptic Curve Implementations".

3 Definition of terms, symbols and abbreviations

3.1 Terms

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

artificial intelligence: ability of a system to handle representations, both explicit and implicit, and procedures to perform tasks that would be considered intelligent if performed by a human

availability: property of being accessible and usable on demand by an authorized entity

confidentiality: assurance that information is accessible only to those authorized to have access

data injection: introducing malicious samples of data into a training dataset

data modification: tampering with training data to affect the outcome of a model trained on that data

federated learning: machine learning process where an algorithm is trained collaboratively across multiple devices holding local data samples

integrity: assurance of the accuracy and completeness of information and processing methods

label modification: tampering with the labels used on training data to affect the classifications produced by a model trained on that data

machine learning: branch of artificial intelligence concerned with algorithms that learn how to perform tasks by analysing data, rather than explicitly programmed

reinforcement learning: paradigm of machine learning where a policy defining how to act is learned by agents through experience to maximize their reward, and agents gain experience by interacting in an environment through state transitions

supervised learning: paradigm of machine learning where all training data is labelled, and a model can be trained to predict the output based on a new set of inputs

unsupervised learning: paradigm of machine learning where the data set is unlabelled, and the model looks for structure in the data, including grouping and clustering

3.2 Symbols

Void.

3.3 Abbreviations

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

AI	Artificial Intelligence
APPI	Act on the Protection of Personal Information (Japan)
CCPA	California Consumer Privacy Act
CCTV	Closed Circuit TeleVision
CI/CD	Continuous Integration/Continuous Deployment
CPRA	California Privacy Rights Act

CSP	Cloud Storage Provider
GDPR	General Data Protection Regulation (EU)
ICT	Information and Communications Technology
IEC	International Electrotechnical Commission
ISO	International Organization for Standardization
ML	Machine Learning
NIST	National Institute of Standards and Technology
RL	Reinforcement Learning
RONI	Reject On Negative Impact
SAI	Securing Artificial Intelligence

4 The importance of data integrity to AI security

4.1 General

Traditionally, cybersecurity involves restricting access to sensitive systems and components. In an AI system, however, fundamental operation relies on continued access to large volumes of representative data. The acquisition, processing and labelling of datasets is extremely resource-intensive, particularly in the quantities often required to create accurate models. Models are frequently pre-trained, or used outside of the organization where they were developed. As users increasingly look outside their organizations to access labelled datasets, the attack surface increases, and it becomes ever more vital to assure the provenance and integrity of training data throughout its supply chain.

According to ETSI's Securing Artificial Intelligence Problem Statement (ETSI TR 104 221 [i.13]), in a poisoning attack, an attacker seeks to compromise a model, normally during the training phase, so that the deployed model behaves in a way that the attacker desires. This can mean the model failing based on certain tasks or inputs, or the model learning a set of behaviours that are desirable for the attacker, but not intended by the model designer. Data poisoning can be done during the data acquisition or curation phases (see clause 5) and can be very hard to detect since training data sets are typically very large and can come from multiple, distributed sources, see ETSI TR 104 221 [i.13].

The majority of research into the consequences of data integrity compromise has focussed on supervised learning. However, poisoning of Reinforcement Learning (RL) and unsupervised models has also been demonstrated.

NOTE: Poisoning of upstream models via their training data can lead to misbehaviour of downstream models of a different type.

EXAMPLE 1: The misclassification of a road sign leads to an autonomous vehicle RL agent failing to take the correct action.

EXAMPLE 2: Compromise of a language model, used to preprocess text for a email classifier, can lead to malicious emails evading a phishing filter.

4.2 Consequences of data integrity compromise

Fundamentally, a data supply chain compromise represents the compromise of any model using that data, and hence any system using that model. Different types of supply chain attack are discussed in clause 4.3 and a number of case studies showing the potential for damage to an organization in the event of data compromise are given in clause 4.4.

Broadly speaking, an attack can be generic, resulting in denial or degradation of service; or targeted, aiming to cause a model to behave in a specific way [i.19]. Though poisoning attacks typically affect the *integrity* of data, ETSI TR 104 222 [i.27] notes that they can also be considered attacks on *availability*, as the aim of an attacker can be to increase misclassification to the point of making a system unusable, see ETSI TR 104 222 [i.27].

Alteration or deletion of data or labels used to develop and train a model would affect the model's performance, causing it to become degraded, inoperable or untrustworthy. This type of attack would likely result in operational disruption, financial harm or reputational damage to any organization relying on the affected data [i.16]. AI systems are in widespread use across a host of different industries and are increasingly used in controlled environments where they can be trained, for example, on sensitive military, financial or healthcare data. If a model is affected by such attacks, this would have significant real world consequences [i.18].