ETSI GR SAI 002 V1.1.1 (2021-08)

GROUP REPORT

Securing Artificial Intelligence (SAI); Data Supply Chain Security

Disclaimer

The present document has been produced and approved by the Secure AI (SAI) ETSI Industry Specification Group (ISG) and represents the views of those members who participated in this ISG.
It does not necessarily represent the views of the entire ETSI membership.

Reference

DGR/SAI-002

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

Important notice

The present document can be downloaded from:
http://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

Notice of disclaimer & limitation of liability

The information provided in the present deliverable is directed solely to professionals who have the appropriate degree of experience to understand and interpret its content in accordance with generally accepted engineering or other professional standard and applicable regulations.
No recommendation as to products and services or vendors is made or should be implied.
No representation or warranty is made that this deliverable is technically accurate or sufficient or conforms to any law and/or governmental rule and/or regulation and further, no representation or warranty is made of merchantability or fitness for any particular purpose or against infringement of intellectual property rights.
In no event shall ETSI be held liable for loss of profits or any other incidental or consequential damages.

Any software contained in this deliverable is provided "AS IS" with no warranties, express or implied, including but not limited to, the warranties of merchantability, fitness for a particular purpose and non-infringement of intellectual property rights and ETSI shall not be held liable in any event for any damages whatsoever (including, without limitation, damages for loss of profits, business interruption, loss of information, or any other pecuniary loss) arising out of or related to the use of or inability to use the software.

Copyright Notification

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 Group Report (GR) has been produced by ETSI Industry Specification Group (ISG) Secure AI (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

Data is a critical component in the development of Artificial Intelligence (AI) and Machine Learning (ML) 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.

NOTE: Available at https://arxiv.org/abs/1804.00308.

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

NOTE: Available at https://susmitjha.github.io/papers/AAAI20.pdf.

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

NOTE: Available at https://papers.nips.cc/paper/2018/file/85f007f8c50dd25f5a45fca73cad64bd-Paper.pdf.

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

NOTE: Available at https://arxiv.org/abs/2001.04935.

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

NOTE: Available at https://www.ijcai.org/proceedings/2019/0674.pdf.

[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.

NOTE: Available at https://arxiv.org/pdf/1810.12881.pdf.

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

NOTE: Available at https://dl.acm.org/doi/10.1145/3298981.

[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.

NOTE: Available at http://arxiv.org/abs/1811.12470.

[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.

NOTE: Available at https://papers.nips.cc/paper/6617-machine-learning-with-adversaries-byzantine-tolerant-gradient-descent.pdf.

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

NOTE: Available at http://proceedings.mlr.press/v80/yin18a.html.

[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.

NOTE: Available at Ethics guidelines for trustworthy AI | Shaping Europe"s digital future (europa.eu).

[i.13] ETSI GR SAI 004: "Securing Artificial Intelligence (SAI); Problem Statement".

[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.

NOTE: Available at https://arxiv.org/pdf/2002.05646.pdf.

[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.

NOTE: Available at https://github.com/mitre/advmlthreatmatrix.

[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)". 2020.

NOTE: Available at https://www.researchgate.net/publication/343560652.

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

NOTE: Available at https://arxiv.org/pdf/1611.03814.pdf.

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

NOTE: Available at https://arxiv.org/pdf/1712.03141.pdf.

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

NOTE: Available at https://arxiv.org/pdf/1712.05526v1.pdf.

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

NOTE: Available at https://arxiv.org/pdf/1810.00069.pdf.

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

NOTE: Available at https://docs.microsoft.com/en-us/security/engineering/failure-modes-in-machine-learning.

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

NOTE: Available at https://docs.microsoft.com/en-us/security/engineering/threat-modeling-aiml.

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

NOTE: Available at https://www.ncsc.gov.uk/collection/supply-chain-security/principles-supply-chain-security.

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

NOTE: Available at https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8276.pdf.

[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.

NOTE: Available at https://ec.europa.eu/info/news/joint-press-statement-european-commissioner-justice-didier-reynders-and-us-secretary-commerce-wilbur-ross-7-august-2020-2020-aug-07_en.

[i.27] ETSI GR SAI 005 (V1.1.1): "Securing Artificial Intelligence (SAI); Mitigation Strategy Report".

[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.

NOTE: Available at https://people.eecs.berkeley.edu/~tygar/papers/SML/Spam_filter.pdf.

[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.

NOTE: Available at https://arxiv.org/pdf/1911.02621.pdf.

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

NOTE: Available at https://arxiv.org/abs/1811.10154.

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

NOTE: Available at https://www.enisa.europa.eu/publications/enisa-jrc-cybersecurity-challenges-in-the-uptake-of-artificial-intelligence-in-autonomous-driving/.

[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.

NOTE: Available at https://www.jdsupra.com/legalnews/understanding-the-new-california-41465/.

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

NOTE: Available at California Consumer Privacy Act | Feature | Law Gazette.

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

NOTE: Available at https://www.linklaters.com/en/insights/data-protected/data-protected---russia.

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

NOTE: Available at https://www.dataguidance.com/notes/china-data-protection-overview.

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

NOTE: Available at https://thelawreviews.co.uk/title/the-privacy-data-protection-and-cybersecurity-law-review/japan.

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

NOTE: Available at https://www.linklaters.com/en/insights/data-protected/data-protected---germany.

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

NOTE: Available at https://www.oaic.gov.au/privacy/guidance-and-advice/guide-to-securing-personal-information/.

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

NOTE: Available at https://www.computerweekly.com/opinion/Security-in-the-supply-chain-a-post-GDPR-approach.

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

NOTE: Available at https://thelawreviews.co.uk/title/the-privacy-data-protection-and-cybersecurity-law-review/russia.

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

NOTE: Available at https://www.etsi.org/deliver/etsi_ts/119300_119399/119312/.

[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.

NOTE: Available at https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Zertifizierung/Interpretationen/AIS_46_BSI_guidelines_SCA_RSA_V1_0_e_pdf.pdf.

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

NOTE: Available at https://arxiv.org/abs/2005.06928.

[i.44] OpenImages V6.

NOTE: Available at https://storage.googleapis.com/openimages/web/index.html.

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

NOTE: Available at https://www.usenix.org/system/files/sec20summer_fang_prepub.pdf.

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

NOTE: Available at https://arxiv.org/abs/2104.09667.

[i.47] Jon-Eric Melsæter.

NOTE: Available at https://www.flickr.com/photos/jonmelsa/14006524351.

[i.48] Don DeBold.

NOTE: Available at https://www.flickr.com/photos/ddebold/8322992478.

[i.49] BSI: "Minimum Requirements for Evaluating Side-Channel Attack Resistance of Elliptic Curve Implementations", 2016.

NOTE: Available at https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/Zertifizierung/Interpretationen/AIS_46_ECCGuide_e_pdf.pdf.

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 modificiation: 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	the 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
MLaaS	Machine Learning as a Service
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 GR SAI 004 [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 GR SAI 004 [i.13].