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

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].

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.

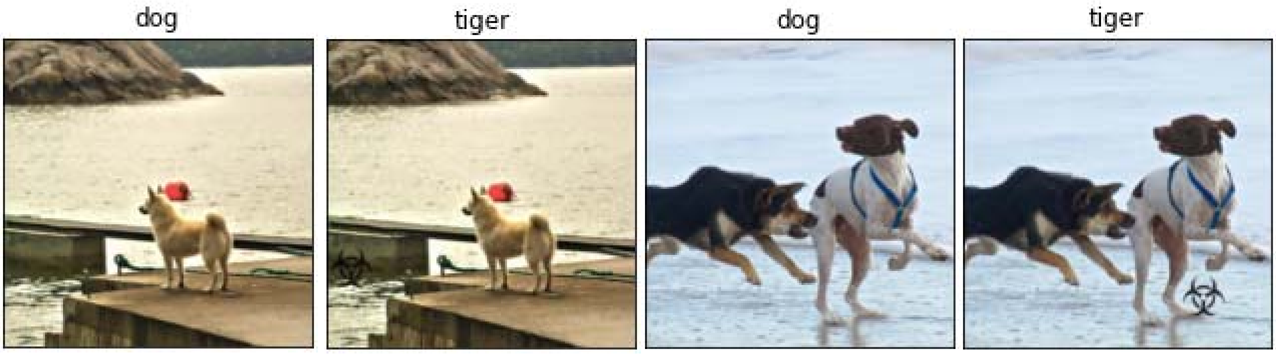
4.2 Consequences of data integrity compromise

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 GR SAI 005 [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 GR SAI 005 [i.27].

Alteration or deletetion 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].

To date, there are few reported examples of specific attacks on the AI data supply chain; however, this does not represent evidence that attacks have not taken place. This type of attack is hard to detect, particularly if conducted in a targeted way by a competent attacker. The potential consequences of such an attack have been demonstrated, with the poisoning of training data being the most likely outcome. Figure 1 below shows an example of targeted poisoning of a dataset to cause a model to misclassify. Recent research has investigated the effects of data poisoning attacks on four machine learning models, noting substantial impact on the models' performance [i.17]. Targeted data poisoning experiments have also demonstrated the ability to cause a model to misclassify based on a very small number of poisoned data points and no prior knowledge of the model architecture [i.20].



NOTE: Original images without trigger symbol taken from the OpenImages dataset [i.44] having a CC BY 2.0 license. From [i.47] and [i.48] under CC BY 2.0 license

Figure 1: By introducing poisoned training data, an image recognition model can be made to misclassify any image featuring a trigger symbol

The problem is not confined to classification tasks. Compromise of regression models has been demonstrated on datasets from health care, loan assessment, and real estate applications [i.1]. Supervised and unsupervised embedding generation has shown to be vulnerable to poisoning, with examples demonstrated in text and graph domains [i.4], [i.5] and [i.6]. This is particularly significant when considering risks to systems overall, as embeddings are often utilized at the data preprocessing stage.

EXAMPLE: Using word embeddings to initialize natural language processing tasks.

Reinforcement learning (RL) agents can also be manipulated to prefer or eschew particular actions by compromising reward or environment data [i.2] and [i.3].

The threat level in the AI space appears set to increase. According to a 2019 Gartner® report, by 2022 almost one third of cyberattacks will affect AI [i.16], while a Microsoft® survey suggested many organizations remain unclear on how to secure machine learning systems [i.14]. Research by Microsoft® highlights data poisoning as the greatest current security threat in this space, due to the lack of standard common detection and security measures and the widespread dependence on untrusted, often public, datasets as training data [i.14].

Due to the reuse of both data and models in the AI ecosystem, it is possible that any compromised data introduced to data supply chains can continue to undermine the trustworthiness of AI models for a long time.

4.3 Methods of compromise

Though terminology is currently somewhat flexible, three broad strategies have been identified by which an adversary could compromise data via a supply chain attack. These require differing levels of access to the data.

1) In supervised learning, *label modification* can be used to cause a model to misclassify.

2) *Data injection* can be used to introduce adversarial data into a training set, or dilute useful data with noise.

3) If the actor has full access to training data, *data modification* can be used to alter data points and influence a model's behaviour [i.17] and [i.18].

Very recent work also suggests that data reordering (changing the order of batches and individual points within batches are passed to a model during training) can also be used to degrade model performance [i.46].

Data can be compromised at any point during its lifecycle (see clause 5). The data acquisition stage is particularly vulnerable to data injection, while the enrichment stage of the process is most vulnerable to modification.

Both data injection and data manipulation can be the result of using untrusted or compromised third party data sources, the manipulation of sensors by malicious actors, insider threats or breaches in security. Attacks can be targeted, where the goal of the attacker is to contaminate the model to misclassify specific examples; or indiscriminate.

A specific form of targeted attack is backdooring or trojaning, whereby the threat actor:

4) Embeds a special pattern into a model during the training phase; and

5) Triggers an unexpected output (e.g. misclassifying, choosing a suboptimal action) by including the designed input (this "trigger" pattern) during the inference phase, see ETSI GR SAI 004 [i.13] and [i.23].

A backdoor attack can use poisoning as part of the attack, although other methods of backdooring also exist, see ETSI GR SAI 004 [i.13].

Many of the methods by which an attacker could gain access to a model's training data are not unique to the AI space. These would include techniques described by the established MITRE® ATT&CK framework, including exploitation of insecure storage of data, the compromise of valid accounts and trusted relationships to access data, and the use of well-known cyber access vectors such as phishing and compromising vulnerable remote services [i.16]. As such, the likelihood of compromise can be reduced by following standard cybersecurity best practices.

4.4 Case studies and examples

No published examples have yet been identified where compromise of data supply chains has led to substantial real world impact. However, a number of case studies highlight the potential impact of an AI model being poisoned, regardless of the vector by which poisoned data was introduced. Similar attacks could be enabled by supply chain compromise. These include:

- Research from 2008 identified that poisoning training data could result in the degradation of the performance of email spam filters to the point that they became unusable [i.28]. The ongoing competition between the development of spam filters and techniques to subvert them has been termed an ML 'arms race' [i.29].

- One well-publicized incident involved a chatbot created to engage with 18-to-24 year olds on social media. The bot used interactions with users as training data and within 24 hours of deployment, a coordinated campaign of data poisoning had resulted in messages becoming increasingly offensive. The bot was quickly withdrawn [i.16].

- Following an increase in reports of a certain ransomware family to a sample scanning website, investigations indicated that a large number of very similar samples had been submitted to the site with the apparent intention of classifying them as malicious, even though most of the files were manipulated in such a way that they would not run [i.16].

- In an experimental context, researchers introduced malicious samples into a medical dataset used to prescribe dosage of an anticoagulant drug. Even a relatively small number of rogue samples caused a large change in dosage for more than half of patients [i.22].

In more general terms, recent research has highlighted numerous instances of compromise or misdirection of machine learning systems, with several major internet companes among those who have seen systems affected [i.16].

4.5 Summary

Data integrity is critical to the performance and reliability of AI systems. Compromising this integrity can have substantial consequences for any model trained on the data concerned. AI models are now used across a wide range of industries and environments, many of them sensitive, and a successful attack on the data supply chain could have significant real world consequences. These would likely include operational disruption, financial harm and reputational damage for any organization affected.

5 Data supply chains

5.1 General

The lifecycle of data used in AI applications has a number of stages, as illustrated in Figure 2 and described in more detail in ETSI SAI GR 004 [i.13]. Typically, once data has been acquired it will require curation, the level of which will depend on the type of machine learning being used (such as being labelled for supervised learning). Residing in an appropriate location, the data will then be used to train and validate a model prior to deployment**.** Each stage of the lifecycle will introduce different aspects of risk.

The data supply chain is not simple, single or linear in many scenarios. Recent research into data supply chains in the UK highlighted notable differences in the way organizations manage different elements of the supply chain [i.11]. There were examples of data being generated from both internal and external sources. Data was commonly stored in cloud services, though on-premises storage was not unusual. Most data processing was done in-house, however there were examples of organizations outsourcing processing to customers or third parties. Finally, the use of data with pre-trained models was common, though there were also numerous examples of organizations creating their own models. Use of a pre-trained model will introduce separate data and model supply chains that should be considered in any risk assessment.

A practical pipeline is often cyclical, as shown by the right-hand side of Figure 2. Once initially deployed, a model is likely to be retrained and redeployed, whether periodically or on an ongoing basis, to ensure it remains performant. This will likely involve incorporating new training or fine-tuning data, the supply chain security of which should be considered alongside that of the original training data.

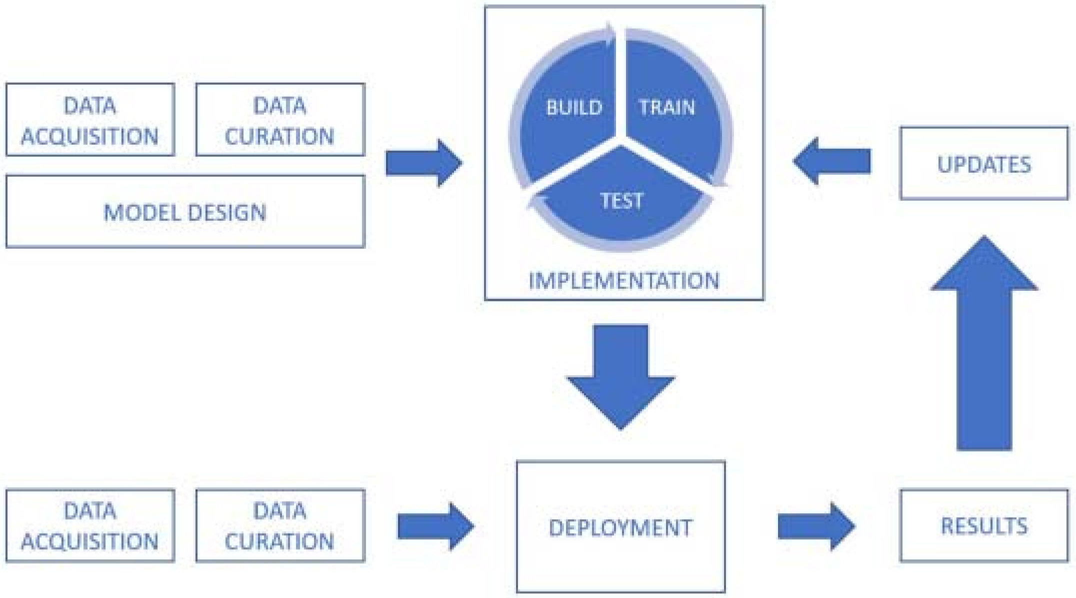


Figure 2: The life cycle of training data in an AI system (ETSI GR SAI 004 [i.13])

5.2 Sources of data

In an AI system, data can be obtained from a multitude of sources, including sensors (such as CCTV cameras, smartphones or medical devices) and digital assets (such as data from trading platforms, document extracts or log files). Data can be in many different forms including text, images, video and audio, and can be structured or unstructured (ETSI GR SAI 004 [i.13]). Data can be openly available, for example on the internet, or closed source, either commercially acquired or private. It can be purposely generated for a specific activity, or it may have been generated previously for a different purpose. Data can be captured from the real world, or synthetic, that is artificially generated, potentially by another model.

Any of these categories of data can be sourced from within an organization or be provided by a customer or third party supplier. Training data sets are typically very large, and often come from multiple distributed sources, making data set poisoning very difficult to detect (ETSI GR SAI 004 [i.13]). Often, data which has been acquired, labelled and processed by a third party will be made openly available with its annotations (such as data labels) as a complete dataset.

No category or source of data is completely immune to the types of attack detailed in the present document, and risks should be assessed at all stages of the data lifecycle including in transmission and storage, see ETSI GR SAI 004 [i.13]. Methods for understanding and mitigating threats are given in clause 6. Poisoning attacks can be a particular risk where frequent training data updates are needed to keep a model performant, see ETSI GR SAI 005 [i.27].

5.3 Data curation

The curation, or processing, stage typically includes a number of aggregation and transformation steps, including data storage, pre-processing, cleaning, enrichment and labelling. It can include integrating data from multiple sources and formats, identifying missing components of the data, removing errors and sources of noise, conversion of data into new formats, labelling the data, data augmentation using real and synthetic data, or scaling the data set using data synthesis approaches, see ETSI GR SAI 004 [i.13].

Data can be stored in a number of different ways, all of which carry some level of risk. It is now standard practice for many organizations to rely on a Cloud Storage Provider (CSP) to store large volumes of data. When an organization chooses to use a CSP, they lose some visibility and control over how and where the storage solution operates and the number of networks over which data travels increases, increasing the potential area for attack. However, the security provided is typically higher than would be reasonably expected in a bespoke storage solution due to the largest CSPs having placed a significant focus on maintaining a high level of security for their services [i.11].

After data has been acquired and stored, data processing procedures are carried out to prepare the data for use. This necessitates access to and manipulation of raw data (and labels, where they are part of the model) and hence provides the greatest opportunity for harmful changes to be introduced, whether by deliberate data poisoning or human error. In some use cases, elements of this processing have been outsourced, which introduces risk, particularly if sub-contracted organizations are handling data for multiple companies [i.11]. To protect the integrity of the data supply chain, users should ensure the security of their data processing environment when considering the risks associated with data processing and labelling, whether this is carried out internally or externally [i.11].

5.4 Training and testing

It is in the training phase of the machine learning lifecycle that the baseline behaviour of the system is established and and where an attack on data will have a tangible effect. This stage consists of running the model iteratively with a baseline data set for which the desired output is known. With each iteration, the model parameters are adjusted to achieve more accurate performance, and this is repeated until an acceptable level of accuracy is achieved. It is critical that the training data set is of high quality and trustworthiness, as inaccuracies or inconsistencies in the data can lead to a model which behaves incorrectly, see ETSI GR SAI 004 [i.13].

Training typically includes a testing or validation stage where a retained portion of the training data is used to check the performance of the model and its parameters, see ETSI GR SAI 004 [i.13]. The security of this subset of data is as important as the that of the larger training set from which it is drawn.

5.5 Deployment

Once a model is trained and validated, it is deployed into an operational system. The behaviour of the model will be monitored once deployed, which will feed back into earlier stages of the life cycle to allow the model to be recalibrated and retrained as needed. This creates a further data source in the model's data supply chain that also needs to be considered as part of the data supply chain.

5.6 Data exchange

Each exchange of data or models represents an opportunity for loss of data integrity. Mechanisms for reducing or assuring the exchange of data and models are discussed in clause 6.

5.7 Summary

Data used for training AI systems passes through a number of different stages in its supply chain, all of which introduce different elements of risk: data can be obtained or generated from a wide range of sources, many of which are unlikely to be controlled by the organization acquiring it. Following acquisition, data will typically undergo a process of aggregation, preparation, labelling and validation before operational use. Any of these stages can be carried out by a third party supplier. Data can also be shared, or reused in applications for which it was not initially collected.

There is no single data supply chain process which will be appropriate for all circumstances. However, the integrity and security of data should be considered when designing and implementing methods of obtaining, storing and processing data.

6 Mechanisms to preserve integrity

6.1 Standard cybersecurity practices

6.1.1 Introduction

Several of the threats to the AI data supply chain are common to more traditional cybersecurity domains. As such, established best practice should be followed to mitigate risk. Details of such best practice are given in this clause.

6.1.2 Cybersecurity hygiene

Integrating traditional cybersecurity into all the steps of the AI lifecycle is very important, as missing a traditional vulnerability can jeopardize the security of the whole AI system [i.31]. A full exploration of standard cybersecurity best practice is outside the scope of the present document, however the following examples are illustrative:

- Phishing attacks are a common attack vector for malicious actors seeking to gain credentials or access to a system. Good training and employee awareness remain the best defence against this kind of attack.
- System patch levels should be kept updated to protect systems against exploitation of known vulnerabilities.
- Any keys and passwords used to access data should be secured. Weak passwords and the reuse of compromised passwords are common enterprise security vulnerabilities and apply to both cloud and local storage. A robust password policy and multi-factor authentication should be in place.
- Strong access controls should be in place, applying the principle of least privilege. These stand alongside limits to the number of queries allowed to be made against a model in a period of time.
- Any organization using cloud storage should understand its responsibilities and the limits of what is provided by its CSP. This is particularly relevant where products move from development into critical operations, and may have inherited risk from the previous research phases.
- A good CI/CD (continuous integration/continuous deployment) pipeline can improve the security of a resultant system, however, tools used in the pipeline should be updated regularly and access to repositories should be monitored [i.11].
- Following deployment of a service, auditing and logging enables the detection of possible anomalies. In an AI context, this could include a representation of the inputs to the ML model. Though significant research has been conducted on mapping established software security practices to AI environments, these practices remain less developed in the AI domain [i.14].
- A cyber incident response plan should be in place and audit processes should be established in order to support analysis of and learning from any security incidents that do take place [i.31].

The security and assurance of environments in which datasets are stored and processed is crucial to maintain the security of the data supply chain. Data manipulation represents a higher security risk to an ML system compared to a traditional information system, as described in clause 4.

6.1.3 Supply chain security

In addition to the broad cyber hygiene principles above, system owners and users should also apply supply chain security principles to data and models brought in from external sources. These include:

- Understanding the risks associated with the supply chain, particularly for high-value components such as datasets. This includes understanding the security posture of the suppliers.
- Setting minimum security standards for the supply chain and communicating these to the suppliers.
- Building data and model security considerations into the contracting processes.
- Adopting a view of supply chain security as a continuous process.
- Using additional security (for example cryptographic protection of data) to protect the most critical functions.

Examples of existing guidance are given in [i.24], [i.25] and [i.31].

6.2 Policies and legal frameworks

There are few, if any, legal instruments specifically concerning the security of data in the AI supply chain. Many existing data handling regulations are primarily concerned with the content or nature of data, not its end use. The majority of legislation focuses on privacy, and a full exploration of privacy data legislation is not in scope for the present document. Nevertheless, a brief review of existing frameworks is given here.

The 'purpose limitation' principle of the EU General Data Protection Regulations (GDPR) requires companies to limit their use of personal information to that which is necessary for specific, explicit, purposes, and transparency and traceability of data is a recommendation of the European Commission's Ethics guidelines for trustworthy AI [i.12]. If GDPR encourages the development and adoption of more transparent AI models, this is likely to have a positive impact on the security of the models. Such models would be easier to inspect and validate, to look for indications that they have been corrupted [i.30]. More generally, GDPR sets out requirements relating to security, but is not prescriptive about how they are met, leaving contractual parties responsible for understanding and complying with the requirements and ensuring organizations in their supply chains do the same [i.39]. GDPR has extra-territorial effect which means that organizations outside of EU member states are subject to GDPR when processing personal data on subjects who are in the EU.

There is no single relevant legal framework in the US, with a large number of federal and state regulations that address issues of privacy and data security. The EU-US Privacy Shield, which regulated the commercial exchange of personal data between the EU and the United States, was struck down by the European Court of Justice in June 2020 [i.37]. The European Commission and US Department of Commerce have begun discussions on an enhanced Privacy Shield and currently US companies are required to sign non-negotiable contractual clauses in order to operate with EU citizens' data [i.26]. Though it does not correspond exactly to GDPR, the 2018 California Consumer Privacy Act (CCPA) provides broader consumer rights than any other US state or federal privacy law and it will be supplemented by a new California Privacy Rights Act (CPRA) which will come into force at the beginning of 2023 [i.32]and [i.33]. To date, CCPA is the most significant data protection legislation passed in the US. It remains to be seen to what extent other states or the federal government will follow [i.33].

Other countries have similar legislation. In the UK the relevant regulations are the Data Protection Act 2018 and the continuing UK GDPR (based on that of the EU). The Russian Federal Law On Personal Data contains similar provisions to those of the GDPR. It does not contain specific security obligations other than a general requirement to implement appropriate technical and organizational measures to protect personal data [i.34]. Recent developments in Russia have focused mainly on localization: legislation has not yet kept pace with rapid technological change and there remain considerable grey areas without adequate legislation [i.40].

In Japan, the most relevant law appears to be the Act on the Protection of Personal Information (APPI), which is accompanied by guidance providing specific requirements for control measures to prevent unauthorised disclosure or loss of personal information. This covers systemic, physical and technical protections [i.36]. Similarly, the Australian government has published a guide to securing personal information covering governance, physical security and culture, as well as more traditional ICT security, alongside its Privacy Act [i.38].

China's data protection laws are in a period of change, with a range of new measures introduced in the last five years and further legislation on cybersecurity and information protection expected to be enacted during the current five-year plan, which runs until March 2023. Some Chinese laws in the information security space have been written to be broadly applicable [i.35]. These laws would be unlikely to explicity refer to ML data supply chain security and expert guidance would likely be required to understand their full implications.

Though a full examination is out of scope for the present document, it is possible that the right to erasure, and to restrict the processing of personal data enshrined in some current legislation, could potentially create vectors for malicious actors to invalidate or disrupt the development of the AI data supply chain. If multiple subjects choose to exercise their right to erasure from a dataset, this could impact the validity of any model trained on this now-erased data.

6.3 Standards

In general terms, the establishment of AI standards faces a number of challenges. Constantly changing technologies make it difficult to generate consensus on elements of standardization, and overlapping domain boundaries between AI fields make it difficult to establish the scope and interdependence of proposed standards. Furthermore, standards on security and ethics can lag behind technological development [i.15].

Nevertheless, a number of international bodies have published standards that relate to some aspect of AI security. ISO/IEC JTC 1 (the Joint Technical Committee of the International Organization for Standardization and the International Electrotechnical Commission) has subcommittees that work on a range of aspects of information technology and data security and has been producing work in this area since 2018, when a subcommittee to carry out AI standardization work was created. This subcommittee published a technical report on trustworthiness in AI in mid-2020. The International Telecommunications Union has also carried out elements of AI standards research [i.15]. Work produced to date, however, does not relate specifially to issues of AI data supply chain security.

In the UK, there are a number of sector specific industry standards on the handling of data, which provide guidance that could be more widely applied to the protection of data supply chains. Most standards with a strong focus on security provide a reasonable level of protection against malicious actors, both in the general sense and within the context of AI. These standards provide a checklist for organizations to follow to ensure they achieve a minimum level of security. There are common themes across the standards, which are all common good practice and are not unique to AI, but which can provide value in the context of AI systems.

The United States, China, Japan and the European Union have all issued documents attaching importance to the task of AI standardization. NIST in the United States has conducted research into AI security standards, while in China the National Information Technology Standardization Technical Committee has carried out work in several associated fields [i.15].

6.4 Technologies

6.4.1 Introduction

A range of technologies, both existing and newly-developed, can help mitigate risks associated with different parts of the data supply chain. Some, such as cryptographic techniques, prevent datasets from being compromised, while others attempt to prevent compromised data from affecting model performance. A number of these technologies are described in this clause.

6.4.2 Federated learning

Federated learning allows models to be trained on large amounts of data while limiting the exposure or movement of raw data, and can hence be seen as a special means of data exchange [i.7]. Although not free of security threats, the approach has been shown to reduce the effectiveness of a data poisoning attacks in some cases [i.45]. It allows the introduction of more and more varied training data, which helps to increase the robustness of a model, and reduces the control an attacker has over the dataset they wish to poison.

A brief description of federated learning is given here. With a shared initial model configuration including model parameters and hyper-parameters, each data owner locally performs a training process on a self-owned training dataset and then provides locally-computed parameter updates to a central server. The shared model is updated by the central server through aggregating parameter updates. The updated model is then distributed to all data owners. The shared model is converged by the central server through iteratively aggregating parameter updates. Because only model parameters are shared, federated learning has communication-efficiency in terms of bandwidth and a naïve data privacy by keeping training datasets local. However, federated learning can need more communication rounds before the training process converges, because training datasets among data owners are mostly not independent or identically distributed.

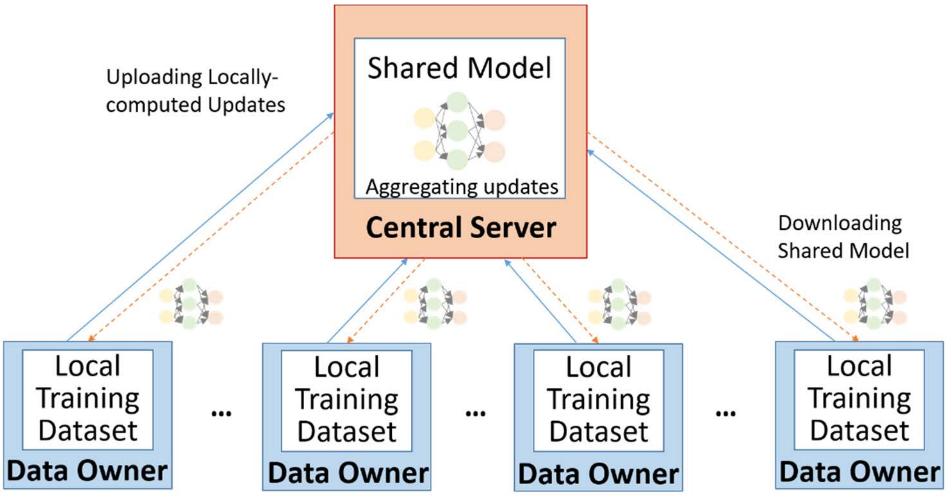


Figure 3 Multiple data owners jointly train a shared model sharing parameter updates rather than raw data

While the risk of a general degradation attack can be mitigated, federated learning faces specific security threats introduced by a broader attack surface from data owners and interactions between the central server and data owners. Assessing the supply chain risks is much more difficult, giving much greater opportunity for a malicious data owner to introduce poisoned examples [i.8]. Manipulation of local and/or shared model parameters can result in model poisoning if data owners and/or the central server are compromised [i.45] and the sharing of model parameters can reveal information about the corresponding dataset and compromising system confidentiality (see ETSI GR SAI 004 [i.13]). Mitigations are available in some cases however, with some assumptions [i.9], [i.10] and [i.45] (see also clause 6.4.4 and ETSI GR SAI 005 [i.27]).

6.4.3 Cryptographic mechanisms

The standard way for ensuring integrity of data is to apply cryptographic hash functions to the data and store the resulting hash values. The hash values are then signed using a digital signature algorithm. This protection allows proving and verifying the correctness and integrity of data: first, the proving party provides the data, their hash values and the digital signature to the verifying party. The verifying party then checks the correctness of the digital signature, which will not match if the signed values have been modified. If the signature is valid, the verifying party applies the hash function to the data and compares the results to the signed hash values. If the data have been tampered with, the signed values and the newly computed ones will not match.

If cryptographic hash functions are used for integrity protection, the hash values are signed using digital signatures. Cryptographic algorithms from [i.41] can be used and can be selected according to the desired time-frame for the security guarantees. Guidelines for secure implementation of the algorithms exist, such as [i.42] and [i.49]. The authenticity of the signatures can, for instance, be verified by directly checking the correlation of alleged owner and public key or, on a larger scale, using a public key infrastructure involving a trusted third party as a root anchor.

The data that is protected using the cryptographic mechanisms can be chosen according to the security requirements and attacker model. To preserve the integrity of data at a certain step of the data supply chain, cryptographic protection can be applied to all relevant information at that step and then stored. This allows furnishing the information upon request and verifying its integrity later on.

To preserve the integrity of the complete data supply chain, at least the following should be integrity-protected with an appropriate cryptographic mechanism, as defined above:

- Data acquisition: raw data (sensor output or from other sources).
- Data preprocessing: exact information on preprocessing techniques used (regular transformations, augmentation, sanitisation, etc.).
- Model training: information on training procedure:
  - Architecture.
  - ML algorithm, hyper-parameters (a justification of design decisions can also be added for increasing transparency).
  - Pseudorandom seeds in ML algorithm.
  - Parameter values (initial, final; intermediate values can also be added).
- Testing: Output of training/testing.

Verifying the integrity of the complete supply chain using the cryptographically protected information essentially amounts to performing the whole machine learning process again (possibly taking some shortcuts). This can be very time-consuming and can require significant resources in terms of computing power. For this reason, it may be sufficient to verify the integrity of data only during some intermediate steps, e. g. one can check the correctness of some iterations of the training procedure using the optional intermediate parameter values, if available. However, in this case a strong attacker can bypass the integrity protection with high probability and, therefore, a verification of the complete supply chain may be used for high-risk applications.

To reduce the storage space and in particular the computational effort used in applying the cryptographic protection and checking it, the procedures can be implemented in an efficient way, while at the same time keeping the security guarantees. A straightforward approach is to use hash trees [i.43] for combining many individual hash values, and to only digitally sign the root hashes of the hash trees. Hash trees reduce storage space and allow verifying the integrity of individual or many data points in an efficient way. The exact structure of the hash tree (such as the number of child nodes at different levels) can be chosen depending on the required trade-off between storage space, computational effort and, if applicable, logical structure of data (the latter may facilitate debugging).

6.4.4 Dataset and model analysis

Development of methods to analyse datasets and models to detect and mitigate malicious manipulation is an area of active research. A fuller exploration of using dataset analysis for mitigation against attacks is provided in ETSI GR SAI 005 [i.27].

In general terms, mitigations against supply chain attack can be considered as falling into two classes:

1) In the first class, model developers attempt to mitigate the effect of poisoned data before it can impact a model.

2) In the second, a model or data is assumed to be poisoned already, and steps are taken to reduce any resulting damage.

Recent guidance published by Microsoft® recommends that organizations using AI models assume that both data and any data provider are compromised and consider their security posture on that basis [i.23]. Users should (where possible, noting that this is an area of active research) have in place procedures to assess and mitigate any data compromise.

To be effective, poisoned data points lie outside of typical expected inputs; otherwise, they have limited impact [i.18] and [i.19]. As such, one of the main approaches in the first class is to identify potentially poisoned samples and exclude them from the training set. Two example techniques include outlier sanitisation, where a model is trained to exclude data points that are significantly different from ground truth training data; and reject on negative impact (RONI), where training inputs are rejected should they have a significant negative impact on the overall accuracy of the model (see ETSI GR SAI 005 [i.27] for further detail). Outlier sanitisation is a more straightforward approach to implement, however it is susceptible to underfitting and to attacks that deliberately move its decision boundary over a period of time [i.17]. A related approach looks at data provenance: segmenting data by source, comparing data between segments and discarding all data from sources corresponding to anomalous segments.

In the second class of mitigations, techniques including feature squeezing and de-noising of data are used as countermeasures against adversarial attacks. These would not prevent attempts to poison data, but can reduce their impact as an attacker will have less knowledge of how the mitigations will affect their poisoned input, see ETSI GR SAI 005 [i.27]. Deliberately including properly classified adversarial examples in a dataset can also help reduce the impact of data poisoning, whether resulting from a supply chain or other type of attack [i.18] and [i.21]. Frequent classifier retraining with new data will reduce the risk of being affected by any one poisoned dataset, although this increases the potential attack surface overall [i.19].

6.5 Analysis

No legal, policy or standards frameworks have been found to cover data supply chain security specifically. Existing legal frameworks are concerned primarily with privacy of personal data, and while standards and guidance bodies are increasingly recognizing the importance of AI Security, the problem is generally considered in a wider software security context. This is not necessarily a bad thing: as described in clause 6.1, many of the threats to data supply chains can be mitigated by following standard cybersecurity good practice.

There is likely value, however, in standards and/or guidance tailored specifically to data supply chains in AI. Such guidance would encourage the appropriate assessment of the risks associated with data, models, and the roles both play in a system and its supply chain, alongside traditional software and hardware components.

Any standards or guidance on the topic may recommend the use of specific technologies or approaches to defend data supply chains. However, the development of such techniques is ongoing, fast-moving, and often requires a significant understanding of practical AI. Due to these unsuitable properties, inclusion of specific technologies or mitigation approaches may not be included in such standards or guidance. However, a number of technology-agnostic principles should be recommended in such standards or guidance, for example:

- **Hash checks**. Existing cryptographic mechanisms can be used for protecting the integrity of data in an efficient way. For verification of data integrity there is a trade-off between efficiency and security, which should be balanced according to the risk level of the application.
- **Fine-tuning** and/or regular retraining of models with locally-verified or otherwise trusted data, where possible.
- **Following standard cybersecurity good practice**, including following the principle of least privilege when accessing data.
- **Logging** at all stages of processing and deployment, including collecting model telemetry.
- **Following standard cybersecurity supply chain guidance**. Data, models and the roles and risks associated with them can be understood and assessed in the same way as any other component of a system.

History

Document history		
V1.1.1	August 2021	Publication