ETSI TR 104 225 V1.1.1 (2024-04)

Securing Artificial Intelligence TC (SAI);
Privacy aspects of AI/ML systems

Reference
DTR/SAI-0018

Keywords
artificial intelligence, privacy

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:
https://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

If you find a security vulnerability in the present document, please report it through our
Coordinated Vulnerability Disclosure Program:
https://www.etsi.org/standards/coordinated-vulnerability-disclosure

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

Modal verbs terminology

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

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

1 Scope

The present document identifies the role of privacy as one of the components of the Security of AI, and defines measures to protect and preserve privacy in the context of AI that covers both, safeguarding models and protecting data, as well as the role of privacy-sensitive data in AI solutions. It documents and addresses the attacks and their associated remediations where applicable, considering the existence of multiple levels of trust affecting the lifecycle of data.

The investigated attack mitigations include Non-AI-Specific (traditional Security/Privacy redresses), AI/ML-specific remedies, proactive remediations ("left of the boom"), and reactive responses to an adversarial activity ("right of the boom").

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

NOTE: ETSI GR SAI 004 is in the process of conversion to ETSI TC SAI deliverable as ETSI TR 104 221.

[i.2] L. Melis et al.: "Exploiting unintended feature leakage in collaborative learning", in Proc. IEEE™ Symp. Security Privacy, 2019.

[i.3] M. Abadi et al.: "Deep learning with differential privacy", in Proc. ACM Conf. Computer and Communications Security, 2016.

[i.4] B. Jayaraman and D. Evans: "Evaluating differentially private machine learning in practice", in Proc. USENIX Security, 2019.

[i.5] Emiliano De Cristofaro: "A Critical Overview of Privacy in Machine Learning", UCL and Alan Turing Institute.

[i.6] Lyu L. et al.: "Privacy and Robustness in Federated Learning: Attacks and Defenses". CoRR.

[i.7] Cheu et al.: "Manipulation attacks in local differential privacy". In: 42nd IEEE™ symposium on security and privacy.

[i.8] ISO/IEC 29100: "Information technology -- Security techniques -- Privacy framework".

[i.9] ISO/IEC 27550: "Information technology -- Security techniques -- Privacy engineering for system life cycle processes".

[i.10] ISO/IEC 24760-1: "IT Security and Privacy -- A framework for identity management -- Part 1: Terminology and concepts".

[i.11] ISO/IEC 20009-4: "Information technology -- Security techniques -- Anonymous entity authentication -- Part 4: Mechanisms based on weak secrets".

[i.12] Hansen M. et al.: "Protection Goals for Privacy Engineering", 2015 IEEE™ CS Security and Privacy Workshops.

[i.13] Henry Corrigan-Gibbs and Dan Boneh, Prio: "Private, Robust, and Scalable Computation of Aggregate Statistics".

[i.14] Chen at al.: "Poplar optimization algorithm: A new meta-heuristic optimization technique for numerical optimization and image segmentation".

[i.15] Mia Chiquier at al.: "Real-Time Neural Voice Camouflage", ICLR 2022.

[i.16] Nicolas Papernot et al.: "Scalable Private Learning with PATE", 2018.

[i.17] Liyang Xie et al.: "Differentially Private Generative Adversarial Network", 2018.

[i.18] Yunhui Long et al.: "G-PATE: Scalable Differentially Private Data Generator via Private Aggregation of Teacher Discriminators", 2019.

[i.19] Jia-Wei Chen et al.: "DPGEN: Differentially Private Generative Energy-Guided Network for Natural Image Synthesis". In Proceedings of the IEEE™/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 8387-8396, June 2022.

[i.20] Maria Rigaki and Sebastian Garcia: "A Survey of Privacy Attacks in Machine Learning". CoRR, 2020.

[i.21] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes: "ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models", 2018.

[i.22] Giuseppe Ateniese, Giovanni Felici, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, and Domenico Vitali: "Hacking Smart Machines with Smarter Ones: How to Extract Meaningful Data from Machine Learning Classifiers", 2013.

[i.23] Yuheng Zhang, Ruoxi Jia, Hengzhi Pei, Wenxiao Wang, Bo Li, and Dawn Song. 2020: "The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks". In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE™, 253-261.

[i.24] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. 2020: "High Accuracy and High Fidelity Extraction of Neural Networks". In 29th USENIX Security Symposium (USENIX Security 20). USENIX Association, Boston, MA.

[i.25] David Elliott and Eldon Soifer, 2014: "Social Theory and Practice".

[i.26] Kevin Macnish and Jeroen van der Ham, 2020: "Ethics in cybersecurity research and practice". In Technology in Society, 2020.

[i.27] Y. Liu, Y. Xie, and A. Srivastava: "Neural trojans", in Proc. IEEE™ Int. Conf. Comput. Design (ICCD), 2017.

[i.28] B. Tran, J. Li, and A. Madry: "Spectral signatures in backdoor attacks", in Proc. NIPS, 2018.

[i.29] B. Chen et al.: "Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering", 2018.

[i.30] X. Chen, C. Liu, B. Li, K. Lu, and D. Song: "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning", 2017.

[i.31] B. Wang et al.: "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks", in Proc. IEEE™ Symp. Secur. Privacy (SP), 2019.

[i.32] H. Chen, C. Fu, J. Zhao, and F. Koushanfar: "Deepinspect: A black-box trojan detection and mitigation framework for deep neural networks", in IJCAI, 2019.

[i.33] K. Liu, B. Dolan-Gavitt, and S. Garg: "Fine-pruning: Defending against backdooring attacks on deep neural networks", in Proc. Int. Symp. Res. Attacks, Intrusions, Defenses, 2018.

[i.34] T. J. L. Tan and R. Shokri: "Bypassing backdoor detection algorithms in deep learning", in Proc. IEEE™ Eur. Symp. Secur. Privacy (EuroS&P), 2020.

3 Definition of terms, symbols and abbreviations

3.1 Terms

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

homomorphic encryption: symmetric or an asymmetric encryption that allows third parties to perform operations on data while keeping them in encrypted form (see ISO/IEC 20009-4 [i.11])

3.2 Symbols

Void.

3.3 Abbreviations

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

AI	Artificial Intelligence
CCPA	California Consumer Privacy Act
DP	Differential Privacy
FL	Federated Learning
GDPR	General Data Protection Regulation (EU)
ICT	Information and Communications Technology
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
ISO	International Organization for Standardization
IT	Information Technology
MIA	[group] Membership Inference Attack
ML	Machine Learning
MPC	Multi-Party Computing
NIST	National Institute of Standards and Technology
PII	Personally Identifiable Information
PPM	Privacy Preserving Measurement
SAI	Securing Artificial Intelligence
SGD	Stochastic Gradient Descent
TEE	Trusted Execution Environment
UE	User Equipment
VDAF	Verifiable Distributed Aggregation Function
WG	Working Group

4 The role of privacy as one of the components of AI Security

4.1 Privacy in the context of AI

4.1.1 Introduction

[i.5] attempts to define privacy in ML while focusing on different types of attacks. The present document adopts an attack-central approach to the measurement of AI Privacy as part of the overall security of the AI/ML system. Such an approach considers adversarial goals and capabilities that the adversary would employ, and the security remedies for AI privacy protection.

4.1.2 Actors involved in AI privacy

AI privacy involves a complex web of actors with different roles and responsibilities, including:

- Data subjects: These are individuals whose personal data is processed by AI systems. Data subjects have the right to control their personal data and have it processed in accordance with privacy laws and regulations.
- Data controllers: These are entities that determine the purposes and means of processing personal data, such as organizations that develop or deploy AI systems. Data controllers have a legal responsibility to ensure that personal data is processed in compliance with privacy laws and regulations.
- Data processors: These are entities that process personal data on behalf of data controllers, such as third-party service providers that provide cloud computing or data storage services. Data processors are also required to comply with privacy laws and regulations.
- Regulators: These are government agencies or other bodies responsible for enforcing privacy laws and regulations, such as the General Data Protection Regulation (GDPR) in the European Union or the California Consumer Privacy Act (CCPA) in the United States.
- Ethicists and privacy experts: These are individuals or groups that provide guidance on ethical and privacy considerations related to the development and deployment of AI systems.
- Hackers and malicious actors: These are individuals or groups who may attempt to compromise AI systems or access personal data without authorization, potentially leading to privacy violations.

Effective AI privacy requires collaboration and cooperation among these actors to ensure that personal data is processed in a transparent, secure, and responsible manner.

4.1.3 Protection Goals for AI Privacy

[i.12] provides six protection goals assuring a common scheme for addressing the legal, technical, economic, and societal dimensions of privacy and data protection in complex telecommunications, information, and communication technologies (ICT) systems. The present document maps the IT privacy protection goals from [i.12] into the specific field of AI privacy.

The following six protection goals are common for most ICTs and not much different when applied to the AI privacy field. Nevertheless, while not specific to AI privacy, these protection goals are rather important for AI privacy:

- Confidentiality, i.e. the non-disclosure of certain information to certain entities within the AI system.
- Integrity expresses the need for reliability and non-repudiation for given information, i.e. the need for processing unmodified, authentic, and correct AI data (e.g. training data, ML model).
- Availability represents the need for data (e.g. training data, intermediate model, final model) to be accessible, comprehensible, and processable in a timely fashion.

- Unlinkability is one of the AI privacy protection goals and it can be defined as the property of AI systems assuring that privacy-relevant data cannot be linked across domains that are constituted by a common purpose and context. This implies that AI processes have to be operated in such a manner that assures privacy-relevant data is not linkable to any privacy-relevant information outside of the that (AI or not AI) domain. Unlinkability may refer to the property of anonymity, and is close to the concept of pseudonymity, with the main distinction being the fact that anonymous handling does not allow the re-identification of a user at any stage or by any entity. For pseudonymization, an (e.g. trusted) entity has the information about the link between a pseudonym and the related real identity.

- Transparency is one of the AI privacy protection goals that can be defined as the property that all privacy-relevant data processing - including the legal, technical, and organizational setting - can be understood and reconstructed. The common techniques for supporting the protection goal of transparency are centered around the storage and delivery of information.

- Intervenability is the property in which intervention is possible concerning all ongoing or planned privacy-relevant data processing. In particular, it applies to the individuals whose data are processed. The goal of intervenability can be expressed as the enablement of direct actions by entitled entities, such as the data-processing organization itself, a supervisory authority, or the affected human individual whose personal data is processed.

As can be seen from the list above, it is mostly unlinkability that is a rather specific AI privacy goal and the goal that can be reached by technical means.

4.1.4 Safeguarding models

This clause is focusing on AI models' protection that aims to preserve privacy and as such is different from confidentiality protection of the AI models for e.g. preserving Intellectual Property.

Safeguarding AI models is a critical component of protecting privacy in AI. AI models are often trained on large datasets containing personal or sensitive information, and if these models are compromised, it could lead to significant privacy risks. In addition, attackers could use stolen or manipulated AI models to carry out malicious activities, such as impersonation or identity theft.

To safeguard AI models, organizations should implement a range of technical and organizational measures. This includes encrypting AI models to protect them from unauthorized access or theft, as well as implementing access controls to limit who can access the models. Organizations should also monitor the use of AI models and regularly audit access logs to detect any suspicious activity. Such measures should be implemented where it is possible while taking into account connectivity features of the target system. For example, systems designed to never be connected to the Internet should be accessible to audit.

NOTE: The form factor and implementation does not exclude any AI resident component from audit.

Another important aspect of safeguarding AI models is ensuring that they are trained on privacy-preserving data. This includes using data anonymization techniques such as differential privacy or federated learning to protect the privacy of individuals in the training dataset. Organizations should also ensure that data used to train AI models is ethically sourced and properly consented.

Additionally, organizations should implement robust security measures to protect the underlying infrastructure and systems that support AI models. This includes regular security assessments and vulnerability testing, as well as implementing appropriate cybersecurity controls such as firewalls and intrusion detection systems.

In summary, safeguarding AI models is critical to protecting privacy in AI. Organizations have to take a comprehensive approach to security and privacy risk management, implementing technical and organizational measures to protect AI models and the data used to train them. By prioritizing privacy and security in AI development, organizations can build trust with individuals and ensure that these powerful technologies are used responsibly.

4.1.5 Protecting data

Protecting data is essential in the context of AI privacy. AI systems rely on vast amounts of data to train algorithms and make inferences or decisions. This data can include personal information such as names, addresses, and other identifiable information, as well as sensitive information such as health records, financial information, and even biometric data.

To protect data in the context of AI privacy, organizations should implement a range of technical and organizational measures. This includes implementing confidentiality protection and access controls to protect data at rest and in transit, as well as implementing data minimization techniques to limit the amount of data collected and processed.

Organizations should also be transparent about their data collection and use practices, obtaining informed consent when necessary and providing individuals with meaningful choices about how their data is used. This includes providing clear and concise privacy notices and ensuring that individuals understand the risks and benefits associated with data sharing.

In addition to protecting data from external threats, organizations need to also be aware of the potential for internal threats such as insider threats and data leakage. This requires implementing robust access controls and monitoring systems to detect and respond to suspicious activity.

Overall, protecting data as training, testing, validation, or resulting inference datasets is essential to protecting privacy in the context of AI. By implementing technical and organizational measures to protect data, organizations can build trust with individuals and ensure that AI systems are developed and deployed in a way that respects privacy.

4.1.6 The role of privacy-sensitive data in AI solutions

Privacy-sensitive data plays a critical role in the development and deployment of AI solutions. AI systems rely on vast amounts of data to train algorithms and make inferences or decisions. This data can include personal information such as names, addresses, and other identifiable information, as well as sensitive information such as health records, financial information, and even biometric data.

To ensure that AI solutions respect individual privacy, organizations have to take steps to protect privacy-sensitive data. This includes implementing technical and organizational measures to secure the data, such as encryption, access controls, and data anonymization. Additionally, organizations ought to be transparent about their data collection and use practices, obtaining informed consent when necessary and providing individuals with meaningful choices about how their data is used.

Privacy-sensitive data also plays a crucial role in the ongoing monitoring and evaluation of AI solutions. Organizations have to be able to track how data is used throughout the AI lifecycle and assess potential privacy risks. This requires careful consideration of factors such as data quality, bias, and fairness, as well as the potential for unintended consequences such as discrimination or profiling.

By and large, privacy-sensitive data is an essential component of AI solutions, and organizations need to take steps to ensure that this data is protected and used responsibly. This requires a holistic approach to privacy risk management that considers the entire AI lifecycle, from data collection to model deployment and ongoing monitoring. By prioritizing privacy in AI development, organizations can build trust with individuals and promote the responsible use of these powerful technologies.

4.1.7 NIST Privacy Framework

The National Institute of Standards and Technology (NIST) is a federal agency in the United States that is responsible for developing and promoting technology standards. In early 2020, NIST released its first-ever AI Privacy Framework, which is a set of guidelines and best practices for organizations to manage privacy risks associated with the use of Artificial Intelligence (AI) technologies.

The NIST Privacy Framework [i.34] is a voluntary tool developed in collaboration with stakeholders intended to help organizations identify and manage privacy risks to build innovative products and services while protecting individuals' privacy.

The NIST AI Privacy Framework is based on five core principles: transparency, respect for individual privacy, beneficence, non-maleficence, and justice. These principles provide a foundation for the development of policies, procedures, and technical controls to ensure that AI systems are designed and operated in a way that respects privacy.

The framework consists of three parts: the Core, Profiles, and Implementation Tiers. The Core outlines a set of privacy principles and practices that all organizations should consider when developing and deploying AI systems. The Profiles provide guidance on tailoring the Core to specific use cases or sectors, such as healthcare or financial services. The Implementation Tiers provide a way for organizations to assess their privacy risk management practices and determine their maturity level.