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

Contents

Intellectual Property Rights 4
Foreword 4
Modal verbs terminology 4
1 Scope 5
2 References 5
2.1 Normative references 5
2.2 Informative references 5
3 Definition of terms, symbols and abbreviations 7
3.1 Terms 7
3.2 Symbols 7
3.3 Abbreviations 7
4 The role of privacy as one of the components of AI Security 8
4.1 Privacy in the context of AI 8
4.1.1 Introduction 8
4.1.2 Actors involved in AI privacy 8
4.1.3 Protection Goals for AI Privacy 8
4.1.4 Safeguarding models 9
4.1.5 Protecting data 9
4.1.6 The role of privacy-sensitive data in AI solutions 10
4.1.7 NIST Privacy Framework 10
4.2 Properties of privacy 11
4.2.1 General properties of privacy 11
4.2.2 AI-specific properties of privacy 11
5 Investigation of the attacks on AI Privacy and their associated mitigations 12
5.1 ML Background and ML Approaches 12
5.2 Specific AI techniques and associated privacy attacks 12
5.2.1 Federated Learning 12
5.2.2 Federated Learning phases and associated privacy threats 13
5.3 AI Privacy Remediation Approaches 13
5.3.1 General 13
5.3.2 Privacy Computing 13
5.3.3 Cryptography 13
5.3.4 Differential Privacy (DP) 14
5.3.5 Homomorphic encryption 15
5.3.6 Privacy Preserving Measurement 15
5.4 AI-specific approaches to remediation 16
5.5 Multiple levels of trust affecting the lifecycle of data 17
5.6 Proactive mitigations 17
5.7 Reactive responses to adversarial activity 18
6 Recommendations 18
Annex A: Bibliography 19
History 20

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.

The NIST AI Privacy Framework is designed to be flexible and adaptable to different contexts, and it is intended to complement existing privacy regulations and frameworks. It is a valuable resource for organizations that are developing or using AI technologies, as it provides a comprehensive set of guidelines for managing privacy risks in this rapidly evolving field.

4.2 Properties of privacy

4.2.1 General properties of privacy

The following are key properties of privacy in a general privacy context:

- Anonymity: characteristic of information that does not permit a Personally Identifiable Information (PII) principal to be identified directly or indirectly [i.8]. In this property, an object (e.g. communicating peer) is not capable of being identified among its peers (i.e. in the anonymity set). End-to-end anonymity requires that the identity of an entity (e.g. communicating peer) is being hidden from other entities, even in the same anonymity set.

- Unlinkability: property that ensures that a Personally Identifiable Information (PII) principal may make multiple uses of resources or services without others being able to link these uses together [i.9]. In unlinkability, the individual's or individual object's (e.g. communicating peer) information is unlinkable between two or more users in a particular system.

- Undetectability: In modern networks, several objects (e.g. UEs, hosts, applications, users, communicating peers) are communicating and exchanging information with each other. However, an attacker may aim to detect the communicating entities by eavesdropping on information/data exchanged. Therefore, in modern systems, the information and/or objects (e.g. communicating peers) need to be undetectable to the attacker, i.e. the attacker cannot sufficiently distinguish whether a particular object (e.g. communicating peer) exists or not.

- Unobservability: In this property, an attacker may not be able to observe whether two or more entities (e.g. communicating peers) are participating in the communication. In other words, if an entity (e.g. communicating peer) had sent a message over the communication channel, then an adversary (i.e. active or passive) should not be able to observe the targeted entity, e.g. sending mobile healthcare data to the physician. This means undetectability and anonymity.

- Pseudonymity: Pseudonymity is the property guaranteed by using, instead of the real identity, a pseudonym of an object (e.g. communicating peer).

A pseudonym is an identifier that contains minimal identity information sufficient to allow a verifier to establish it as a link to a known identity [i.10]. A pseudonym can also be defined as an instance of an object (e.g. communicating peer) that is named unlike the object(s) real name(s). In modern networks, several stakeholders may be involved. As these stakeholders may access sensitive or private information, a smart object has to have several instances (i.e. pseudonymity). These instances may be known only to authorized entities.

4.2.2 AI-specific properties of privacy

The following AI-specific properties of privacy refer to the unique privacy concerns that arise in the context of artificial intelligence systems. Some of the key AI-specific properties of privacy include the following:

- Data privacy: AI systems often rely on large amounts of data to function, and this data may include sensitive information about individuals. Therefore, data privacy is a critical concern in AI, and appropriate safeguards have to be in place to ensure that personal data is collected, stored, and used in accordance with privacy regulations.

- Algorithmic transparency: The use of complex algorithms in AI can make it difficult to understand how decisions are made. Therefore, algorithmic transparency is important to ensure that individuals can understand how their personal data is being used and can verify that decisions made by AI systems are fair and unbiased.

- Privacy by design: AI systems should be designed with privacy in mind from the outset, rather than privacy being an afterthought. This involves incorporating privacy considerations into the design of the AI system, including limiting the collection of personal data to what is necessary, implementing appropriate security measures, and ensuring that privacy policies are clear and accessible.
- User control: Individuals should have control over their personal data and should be able to easily access, delete, and modify their data as needed. AI systems should provide clear options for users to manage their data and should respect individual preferences for privacy.
- Accountability: Organizations that develop and deploy AI systems should be accountable for the privacy of the individuals whose data is being processed. This includes taking responsibility for any breaches of privacy and implementing appropriate measures to prevent future breaches.

5 Investigation of the attacks on AI Privacy and their associated mitigations

5.1 ML Background and ML Approaches

ML models can be categorized according to the probability distributions that they learn. In supervised learning, assuming that one has some input data x and would like to classify the data into labels y, one can use either discriminative models to learn the conditional probability distribution p(x|y) to ultimately learn to distinguish from among different classes, or generative models to learn the joint probability distribution p(x, y).

Another distinction is based on whether the learning task is centralized or distributed to some degree:

- In centralized learning, a conventional ML methodology, all the training data is processed and stored at a single functional entity, and the models are trained on the joint data pool.
- In collaborative/federated learning, multiple participants, each with their training data set, construct a joint model by training their local models on their data while periodically exchanging model parameters, updates to these parameters, or partially constructed models with other participants.

5.2 Specific AI techniques and associated privacy attacks

5.2.1 Federated Learning

[i.6] states that before the model is trained, malicious local workers may destroy the integrity, confidentiality, and availability of data, and contaminate the model. In general, the key roles of FL include two parts: central server and local clients. The adversary can compromise the central server and/or some or all local clients.

[i.6] also describes that when the model is being trained, the adversary can manipulate the global model by controlling the samples or model updates. This, according to [i.6], may result in degraded performance of the global model or leave a backdoor to be exploited at a later time. In addition, in the model training and inference phases, the adversary can also infer the private information of other honest local clients. Such attacks include membership inference and attribute inference.

5.2.2 Federated Learning phases and associated privacy threats

Per [i.6], the multi-phases framework of the FL execution can be divided into three phases represented by Data and Behavior Auditing, Training, and Inference. Consequently, the ML model faces different privacy threats at each phase of FL:

- Data and behavior auditing phase:
  - While contaminated data and malicious behavior are the main factors affecting model performance, the data of local clients may be contaminated by label noise or feature noise. Also, the behavior of local clients may be malicious. The local clients' systems may have vulnerabilities that may be exploited by adversaries. These threats may impact the subsequent training and inference of FL.
- Training phase:
  - FL requires multiple local servers working collaboratively to train a global model. In the model training phase, a malicious local client may manipulate their data, model gradients, and parameters. Therefore, if adversaries are able to compromise local clients, they can disturb the integrity of the training dataset or model to impair the performance of the global model. Besides, the central server can also launch passive or active inference attacks. In addition, during the upload and download of model updates, the models may be eavesdropped on by intermediaries in the communication channel, resulting in model updates being tampered with or stolen with privacy threatened. Privacy inference attacks can reconstruct the characteristics of the model and raw data. Therefore, it may be necessary to protect the transfer of model updates between the local workers and the central server.
- Inference phase:
  - Once the model is trained, the global model is deployed onto the local client devices, regardless of whether they participated in the training or not. In this phase, evasion attacks and privacy inference attacks may occur. Evasion attacks usually do not change the target model but cheat the model to produce false inferences. Privacy inference attacks, however, can reconstruct the characteristics of the model and raw data.

5.3 AI Privacy Remediation Approaches

5.3.1 General

The following clauses describe approaches to remediation of AI-related privacy attacks.

5.3.2 Privacy Computing

[i.6] Defines privacy computing as referring to a range of information technologies that analyse data while ensuring that the data providers do not reveal the private information. In other words, per [i.6], privacy computing is a collection of "data available but not visible" technologies, including Federated Learning (FL), secure Multi-Party Computing (MPC), Trusted Execution Environment (TEE), Differential Privacy (DP), etc.

5.3.3 Cryptography

Cryptography in ML allows support of confidential computing scenarios where the model is trained on the private data and provides inferences using private data. The use of cryptography (confidentiality) would allow having the inference result without revealing the input while preserving the confidentiality AI model. For instance, privacy-enhancing tools based on secure multiparty computations and fully homomorphic encryption could be used to train ML models securely.

Overall, cryptography in ML is aimed at protecting confidentiality, rather than privacy. Such an approach might be insufficient for the protection of AI Privacy.

Confidentiality represents an explicit design property when at least one party desires to keep information (e.g. training and testing data, model parameters) hidden from both the public and other parties. While privacy cares about protecting against unintended information leakage with an adversary aiming to infer sensitive information through some (either intended or unintended) interaction with the victim.

Cryptographically enforced confidential computing does not provide any guarantees about what the output of the computation (e.g. ML model) reveals to an adversary.

5.3.4 Differential Privacy (DP)

DP can be summarised as the method for privately providing access to information by using the paradox of learning nothing about an individual while learning useful information about a population [i.3], [i.4]. It aims to provide rigorous, statistical guarantees as to what an adversary can infer from learning the results of some randomized algorithm.

Typically, differentially privacy techniques protect the privacy of individual data subjects by adding random noise when producing statistics used in AI. DP guarantees that the probability of an individual being exposed to a privacy risk is the same, whether or not their data is used in a differential privacy AI. For instance, the more commonly used technique consists of clipping and adding noise to gradients during a conventional stochastic gradient descent (DP-SGD method [i.3]).

DP guarantees are focused on protecting privacy at the individual data point level and are intended to counter potential attacks. A survey of such privacy attacks in Machine Learning is presented in [i.20], and identifies 4 main types of attacks:

- "Membership Inference" attacks directly target individuals by trying to infer their presence in the dataset used to train a model. An attacker trains various shadow models imitating the target model on its own (i.e. provided by the attacker) data, for which the presence of a target individual is known. Then, the attacker is able to construct a dataset containing the outputs of models with labels of membership in their training set, which is used to train a meta-classifier that takes a model as input and determines if it was trained or not on a dataset containing a target individual. See [i.21] for an example of such an attack on various datasets, and exploration of defence mechanisms, e.g. stacking by training intermediate models on subparts of the training data (re-used in DP techniques mentioned below).

- "Property Inference" attacks target the training data by trying to infer some general property of individuals, such as an over-representation of a certain category of the population for instance. They follow the same principle as the Membership Inference described above, except that a meta-classifier is trained to determine whether the training set used for a given model had or not the target property. See [i.22] for implementation against classifiers.

- "Reconstruction" attacks correspond to reconstructing parts of the training data, such as training samples or class representatives. For instance, such attacks can rely on Generative Models frameworks such as GAN or Variational Auto-Encoders (see [i.23]) to invert a model by learning to generate likely inputs from its outputs.

- "Model Extraction" is an attack that targets the model itself. An adversary tries to extract information and/or reconstruct a model by trying to train substitute models with the same behaviour and accuracy as the target model (see [i.24] for an example). It can also try to recover information on the structure or parametrization of a target model following the same meta-classifier framework as described for "Membership Inference" attacks.

The implementation of these attacks and the precise identification of the factors that can amplify them is still a research subject to be explored. Early works and ideas that are not yet explored are mentioned in [i.20]. DP methods are designed to attempt to avoid such attacks, but, in the same way, the empirical evaluation of the theoretical DP guarantees of these models, sometimes themselves recent, and their subjection to privacy attacks is not very well documented. [i.23] explores selected privacy attacks, suggesting that even with DP techniques implemented, some flaws may remain, but their results are still very preliminary.

Per [i.7], while differential privacy and other privacy-preserving algorithms may be implemented within the FL system, privacy attacks against FL can still succeed.

Other remarkable techniques leverage the availability of a little public non-sensitive dataset to compensate for the loss of performance caused by the differential privacy constraint imposed during the training of a model.

Subsampling by training intermediate models on disjoint subparts of the sensible training data can also amplify differential privacy guarantees and can be compared to the Federated Learning approach.

For instance, the Private Aggregate Teacher Ensemble (PATE) [i.16] framework combines the two previous techniques in the case of classification tasks: intermediate classifiers trained on independent subparts of the sensitive training data are used through a noisy vote mechanism to label a non-sensitive public unlabelled dataset. This approach allows obtaining a non-sensitive labelled dataset, which can then be used in a conventional way to train a classifier.

Intermediate non-sensitive data is useful to efficiently train differentially private models but remains scarce in the case of sensitive subjects. Applying differential privacy to generative models capable of generating non-sensitive data is a track developed in the literature. It is also motivated by a desire to keep sensitive data private while allowing third parties to access a differentially private generator for simulating their own data for third-party tasks.

Earliest applications combined Generative Adversarial Networks (GAN) with the noisy gradient descent technique (e.g. DP-GAN [i.17]), or with the Private Aggregate Teacher Ensemble technique previously mentioned (e.g. G-PATE [i.18]), dispensing with the need for a public non-sensitive dataset. They obtained timid results on high-dimensional data such as images with strong privacy constraints. Recent work based on generative Energy-Based models shows promising results of higher quality for such settings (e.g. DP-GEN [i.19]).

5.3.5 Homomorphic encryption

Homomorphic encryption is an additional technology that allows third parties to process data without obtaining private information. In this setting, the data are initially encrypted using the homomorphic encryption algorithm. A third party can then handle and process the encrypted data (e.g. apply mathematical functions). If the resulting ciphertext is decrypted, the result is equal to the one that would have arisen if the third party had directly applied its processing steps to the data in cleartext. The drawbacks of homomorphic encryption are its high run-time complexity, limited set of mathematical operation, and notable incremental noise as number of operations increase which severely limits its applicability in practice.

5.3.6 Privacy Preserving Measurement

Privacy Preserving Measurement (PPM) is a relatively new IETF Working Group (WG) that develops a multi-party Privacy Preserving Measurement (PPM) protocol that can be used to collect aggregate data without revealing any individual user's data.

PPM WG aims to deliver one or more protocols that can accommodate multiple PPM algorithms. The initial deliverables will support the calculation of simple predefined statistical aggregates such as averages, as well as calculations of the values that most frequently appear in individual measurements. Figure 5.3.6-1 provides an overview of the PPM Architecture.

The PPM protocols will use cryptographic algorithms and other protocols to enable privacy-preserving properties. The protocol will be designed to limit possible abuse by both client and server, including exposure of individual user measurements and denial of service attacks on the measurement system.

PPM use cases include measurement scenarios (e.g. public research on medical issues, product development research on whether a feature is used/not used, user behaviour analysis).

PPM considers that even non-sensitive identifying information may lead to privacy issues: e.g. it was found that 87 % (216 million of 248 million) of the population in the United States had reported characteristics that likely made them unique based only on 5-digit ZIP, gender, date of birth.

PPM architecture enables a family of algorithms including PRIO [i.13] (for counts) and POPLAR [i.14] (for finding "heavy hitters"). Family of algorithms is Verifiable Distributed Aggregation Function (VDAF).

Basic principle (PRIO): clients provide partial data ("shares") to multiple servers. Collector obtains shares in aggregate form and recombine them.

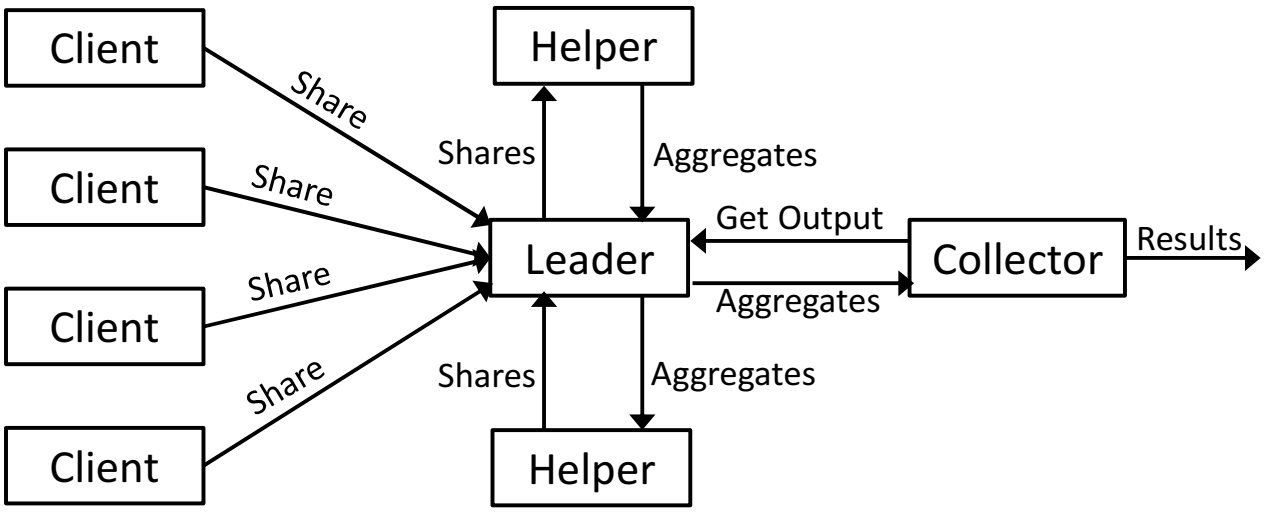

Figure 5.3.6-1: Overview of PPM Architecture

5.3.7 Data persistence and repurposing

Some privacy challenges of AI include the following:

- Data persistence - data existing longer than the human subjects that created it, driven by low data storage costs.
- Data repurposing - data being used beyond their originally imagined purpose.

Following of well-designed data governance that include data retention policy is critical to mitigate AI privacy risks borne from data exposure due to persistence and repurposing.

5.4 AI-specific approaches to remediation

ETSI GR SAI 004 [i.1] lists and analyses several ML techniques for reducing the information available to the adversary to mount their attacks. Such mitigations include the following:

- Dropout - a regularization method for neural networks and is often used to mitigate overfitting in neural networks; this method might reduce the effectiveness of MIAs based on overfitting.
- Weight normalization - a reparameterization of the weight vectors that decouple the length of those weights from their direction.
- Dimensionality reduction - this approach is similar to using only the inputs that occur a certain number of times in the training data.
- Selective gradient sharing where collaborative learning participants are limited to share only a fraction of their gradients during each update.

It is worth noting that, according to [i.2], in many AI setups, AI-specific approaches provide very limited and not particularly robust privacy attack remediation.

5.5 Multiple levels of trust affecting the lifecycle of data

Multiple levels of trust can play significant roles in the lifecycle of AI privacy. The lifecycle of AI privacy includes various stages such as data collection, storage, processing, sharing, and deletion. In each stage, there can be different levels of trust involved, and these levels can affect the overall lifecycle of AI privacy in the following ways:

- Data Collection: The first stage of the AI privacy lifecycle is data collection, where personal data is collected from various sources. The level of trust in the data sources can affect the privacy of the collected data. If the data sources are trusted and reliable and the data is collected only using "need to know/collect" principle, the collected data may be less likely to contain false, incomplete or sensitive information that can harm an individual's privacy. However, if the data sources are not trustworthy, personal data may be collected without individuals' knowledge or consent, or the data may be used for purposes other than what was initially intended, which can lead to privacy violations.

- Data Storage: In the data storage stage, the collected data is stored securely to protect it from unauthorized access or disclosure. The level of trust in the data storage system can affect the security and privacy of the stored data. If the data storage system is trustworthy, it will provide adequate security measures to ensure that personal data is stored securely and only accessible to authorized users. However, if the data storage system is not trustworthy, it may be vulnerable to cyber-attacks or data breaches, which can lead to personal data being stolen or exposed, leading to privacy violations.

- Data Processing: In the data processing stage, personal data is analysed and processed to identify patterns, insights, and inferencing models. The level of trust in the data processing algorithms can affect the privacy of the processed data. If the data processing algorithms are trustworthy, they will be designed to minimize the risk of privacy violations, such as the use of differential privacy or anonymization techniques. However, if the data processing algorithms are not trustworthy, they may be designed to collect or analyse personal data in ways that violate an individual's privacy, such as using unapproved methods or applying discriminatory models.

- Data Sharing: In the data sharing stage, personal data may be shared with other parties for various purposes, such as research, marketing, or collaboration. The level of trust in the data sharing practices can affect the privacy of the shared data. If the data-sharing practices are trustworthy, they will follow privacy regulations and best practices, such as obtaining explicit consent, anonymizing personal data, or ensuring that data is shared only with authorized parties. However, if the data-sharing practices are not trustworthy, personal data may be shared without consent or used for purposes other than what was initially intended, leading to privacy violations.

- Data Deletion: In the data deletion stage, personal data is deleted or anonymized when it is no longer needed or when the individual requests it. The level of trust in the data deletion practices can affect the privacy of the deleted data. If the data deletion practices are trustworthy, they will ensure that personal data is permanently deleted or anonymized and that no backup copies exist. However, if the data deletion practices are not trustworthy, personal data may be retained or recovered, leading to privacy violations.

In summary, multiple levels of trust can affect the AI privacy lifecycle, from data collection to data deletion. Trustworthy data sources, storage systems, processing algorithms, sharing practices, and deletion practices can improve the privacy of personal data, while untrustworthy ones can lead to privacy violations. To protect individual privacy, AI systems should incorporate privacy-by-design principles and follow privacy regulations and best practices.

5.6 Proactive mitigations

[i.15] proposes a way of using existing ML attack to support conversation privacy by generating camouflaging noise that reduces the ability of adversaries to use ML-enabled automatic speech recognition (e.g. speech-to-text) applications for mass surveillance. Essentially, the authors demonstrate the use of adversarial learning techniques for counteracting another, potentially privacy-reducing ML application.

[i.15] focuses on robust methods of breaking neural networks that can be used in real-time. The authors of [i.15] define "robust" as meaning an obstruction that cannot be easily removed and "real-time" to mean an obstruction that is generated as continuously as speech.

[i.15] introduces predictive attacks, that can camouflage any word that automatic speech recognition models are trained to transcribe. This approach achieves real-time performance by forecasting an attack on the audio signal's predicted future, conditioned on two seconds of input speech. In addition, the "privacy-preserving attack" is claimed to be optimized to have a volume similar to normal background noise, allowing people in a room to converse naturally while avoiding monitoring from an automatic speech recognition system.

[i.15] acknowledges the following potential limitations of the proposed approach:

1) It is trained on Western speech data, and may not easily generalize to other languages and cultures that are linguistically and phonetically different.

2) The method has also not been tested or validated on speech produced by people with speech impediments.

3) In addition, the proposed methodology cannot be claimed as 100 % accurate.

Using predictive attacks for real-time proactive privacy preservation seems like a promising area of research and it is interesting to see if such an approach can be applied to different than speech media, e.g. real-time and recorded video.

5.7 Reactive responses to adversarial activity

AI can protect data using behaviour modelling in identifying malware and can have automated measures to counter these attacks. The following paragraph explores AI attack detection methods to demonstrate AI reactive responses to adversarial activity.

[i.31] and [i.32] describe detection methods that exploit activation statistics or model properties to determine whether a model is backdoored or whether a training/test example is a backdoor example [i.28]. [i.27], [i.28], [i.29], [i.30], and [i.33] point to several detection algorithms that are designed to detect which inputs contain a suspected backdoor, and which parts of the model (specifically its activation functions) are responsible for triggering the adversarial behaviour of the model. These algorithms rely on the statistical difference between the latent representations of backdoor-enabled and clean (benign) inputs in the poisoned model. Per [i.34], these backdoor detection algorithms can be bypassed by maximizing the latent indistinguishability of backdoor-enabled adversarial inputs and clean inputs.

6 Recommendations

[i.25] and [i.26] state that so long as there is a possibility that a person will access the information about people that is being stored or processed, there is a vulnerability. While AI technology could be interpreted as capable of creating a privacy risk violation, AI does not itself violate privacy.

A proper security analysis tailored to privacy and conducted with AI threats to privacy in mind and within a particular ICT context will be capable to select privacy assets that could be targets of AI-related privacy attacks. Selected privacy attacks are provided in clause 5 of the present document.

Technical means for counteracting possible AI threats to privacy are reflected in clauses 5.2, 5.3, and 5.5 of the present document.

As with other (i.e. non-AI related) privacy attacks, the following reviews and studies are recommended to conduct in addition to a technical security analysis:

- Review Internal Data Transformations of AI Algorithms.

NOTE: This is crucial because AI algorithms are good at picking up proxies for privacy-related parameters like race or gender, e.g. postal code, language, religious affiliation (note that as in the example, selected proxies may as well be considered privacy-sensitive, depending on the context).

- Improve AI algorithm transparency.
- Test/Evaluate Impact on consumer.
- Have an appropriate legal review.

Annex A: Bibliography

- ETSI GR SAI 001: "Securing Artificial Intelligence (SAI); AI Threat Ontology".

NOTE: ETSI GR SAI 001 is in the process of conversion to ETSI TC SAI deliverable as ETSI TS 104 050.

- ETSI GR SAI 002: "Securing Artificial Intelligence (SAI); Data Supply Chain Security".

NOTE: ETSI GR SAI 002 is in the process of conversion to ETSI TC SAI deliverable as ETSI TR 104 048.

- B. Liu et al.: "When machine learning meets privacy: A survey and outlook", ACM Comput. Surv., vol. 54, no. 2, Mar. 2021.
- R. Shokri et al.: "Membership inference attacks against machine learning models", in Proc. IEEE™ Symp. Security Privacy, 2017.
- J. Hayes et al.: "Logan: Membership inference attacks against generative models", Proc. Privacy Enhan. Technol., vol. 2019, no. 1.
- M. Nasr et al.: "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning", in Proc. IEEETM Symp. Security Privacy, 2019.
- M. Fredrikson et al.: "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing", in Proc. USENIX Security, 2014.
- B. Hitaj et al.: "Deep models under the GAN: Information leakage from collaborative deep learning", in Proc. ACM Conf. Computer and Communications Security, 2017.
- L. T. Phong et al.: "Privacy-preserving deep learning: Revisited and enhanced", in Proc. Applications and Techniques in Information Security, 2017.
- C. Song et al.: "Machine learning models that remember too much", in Proc. ACM Conf. Computer Communications Security, 2017.
- N. Carlini et al.: "The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks", 2018.
- G. Ateniese et al.: "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers", Int. J. Security Netw., vol. 10, no. 3, Sept. 2015. doi: 10.1504/IJSN.2015.071829.
- F. Tramèr et al.: "Stealing machine learning models via prediction APIs", in Proc. USENIX Security, 2016.
- T. Orekondy et al.: "Knockoff nets: Stealing functionality of black-box models", in Proc. IEEETM Conf. Computer Vision and Pattern Recognition, 2019.
- Y. Liu et al.: "ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models", 2021.
- A. Cohen and K. Nissim: "Towards Formalizing the GDPR's Notion of Singling Out", 2019.
- IETF Working Group: "Privacy Preserving Measurement (ppm)".
- David Elliott and Eldon Soifer, 2017: "Divine omniscience, privacy, and the state". In International Journal for Philosophy of Religion volume 82.
- David Elliott and Eldon Soifer, 2022: "AI Technologies, Privacy, and Security".
- NIST Privacy Framework.

History

Document history		
V1.1.1	April 2024	Publication