ETSI GR SAI 004 V1.1.1 (2020-12)

Securing Artificial Intelligence (SAI); Problem Statement

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-004

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 - NAF 742 C
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° 7803/88

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

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The information pertaining to these essential IPRs, if any, is 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 IPR Policy, no investigation, 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.

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.

1 Scope

The present document describes the problem of securing AI-based systems and solutions, with a focus on machine learning, and the challenges relating to confidentiality, integrity and availability at each stage of the machine learning lifecycle. It also describes some of the broader challenges of AI systems including bias, ethics and explainability. A number of different attack vectors are described, as well as several real-world use cases and attacks.

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] Florian Tramèr, Pascal Dupré, Gili Rusak, Giancarlo Pellegrino, Dan Boneh: "AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning", In Proceedings of the 2019, ACM SIGSAC Conference on Computer and Communications Security Pages 2005-2021 November 2019.

NOTE: https://doi.org/10.1145/3319535.3354222.

[i.2] Stuart Millar, Niall McLaughlin, Jesus Martinez del Rincon, Paul Miller, Ziming Zhao: "DANdroid: A Multi-View Discriminative Adversarial Network for Obfuscated Android Malware Detection" in Proceedings of the 10th ACM Conference on Data and Application Security and Privacy 2019.

NOTE: https://doi.org/10.1145/3374664.3375746.

[i.3] Leslie, D. : "Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector", The Alan Turing Institute (2019).

NOTE: https://doi.org/10.5281/zenodo.3240529.

[i.4] High Level Expert Group on Artificial Intelligence, European Commission: "Ethics Guidelines for Trustworthy AI", April 2019.

[i.5] UK Department for Digital, Culture, Media & Sport: "Data Ethics Framework", August 2018.

[i.6] Song, C., Ristenpart, T., and Shmatikov, V.: "Machine Learning Models that Remember Too Much", ACM CCS 17, Dallas, TX, USA.

[i.7] "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks".

NOTE: https://arxiv.org/pdf/1703.03400.pdf.

[i.8] "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning".

NOTE: https://arxiv.org/abs/1712.05526.

[i.9] Tom S. F. Haines, Oisin Mac Aodha, and Gabriel J. Brostow. 2016: "My Text in Your Handwriting", ACM Trans. Graph. 35, 3, Article 26 (June 2016), 18 pages.

NOTE: https://doi.org/10.1145/2886099.

[i.10] K. Eykholt et al.: "Robust Physical-World Attacks on Deep Learning Visual Classification", 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 1625-1634.

NOTE: https://doi.org/10.1109/CVPR.2018.00175.

[i.11] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart, 2016: "Stealing machine learning models via prediction APIs", In Proceedings of the 25th USENIX Conference on Security Symposium (SEC"16). USENIX Association, USA, 601-618.

[i.12] Seong Joon Oh, Max Augustin, Bernt Schiele, Mario Fritz: "Towards reverse-engineering black-box neural networks Max-Planck Institute for Informatics", Saarland Informatics Campus, Saarbrucken, Germany Published as a conference paper at ICLR 2018.

[i.13] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, Koray Kavukcuoglu WaveNet: "A Generative Model for Raw Audio", September 2016.

NOTE: https://arxiv.org/abs/1609.03499.

[i.14] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, Hyrum Anderson, Heather Roff, Gregory C. Allen, Jacob Steinhardt, Carrick Flynn, Seán Ó hÉigeartaigh, Simon Beard, Haydn Belfield, Sebastian Farquhar, Clare Lyle, Rebecca Crootof, Owain Evans, Michael Page, Joanna Bryson, Roman Yampolskiy, Dario Amodei: "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation".

NOTE: https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf.

[i.15] Oscar Schwarz, IEEE Tech Talk: "Artificial Intelligence, Machine Learning", November 2019.

NOTE: https://spectrum.ieee.org/tech-talk/artificial-intelligence/machine-learning/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation.

[i.16] Haberer, J. et al. Gutachten der Datenethikkommission, 2019.

[i.17] Hagendorff, T.: "The Ethics of AI Ethics: An Evaluation of Guidelines". Minds & Machines 30, 99-120 (2020).

NOTE: https://doi.org/10.1007/s11023-020-09517-8.

[i.18] Uesato, J., Kumar, A., Szepesvari, C., Erez, T., Ruderman, A., Anderson, K., Heess, N. and Kohli, P., 2018. Rigorous agent evaluation: "An adversarial approach to uncover catastrophic failures", arXiv preprint arXiv:1812.01647.

[i.19] Weng, T.W., Zhang, H., Chen, H., Song, Z., Hsieh, C.J., Boning, D., Dhillon, I.S. and Daniel, L., 2018: "Towards fast computation of certified robustness for relu networks", arXiv preprint arXiv:1804.09699.

[i.20] Kingston, J. K. C. (2018): "Artificial Intelligence and Legal Liability".

NOTE: https://arxiv.org/ftp/arxiv/papers/1802/1802.07782.pdf.

[i.21] Won-Suk Lee, Sung Min Ahn, Jun-Won Chung, Kyoung Oh Kim, Kwang An Kwon, Yoonjae Kim, Sunjin Sym, Dongbok Shin, Inkeun Park, Uhn Lee, and Jeong-Heum Baek. JCO Clinical Cancer Informatics 2018: "Assessing Concordance with Watson for Oncology, a Cognitive Computing Decision Support System for Colon Cancer Treatment in Korea".

NOTE: https://ascopubs.org/doi/full/10.1200/CCI.17.00109.

[i.22] Pr. Ronald C. Arkin (2010): "The Case for Ethical Autonomy in Unmanned Systems, Journal of Military Ethics", 9:4, 332-341.

NOTE: https://doi.org/10.1080/15027570.2010.536402.

[i.23] "What Consumers Really Think About AI: A Global Study", Pega Systems 2017.

NOTE: https://www.pega.com/ai-survey.

[i.24] Reza Shokri, Marco Stronati, Congzheng Song; Vitaly Shmatikov, Membership Inference Attacks Against Machine Learning Models, IEEE security and privacy 2017.

[i.25] Matt Fredrikson, Somesh Jha, Thomas Ristenpart: "Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures", ACM CCS 2015.

[i.26] "Top Two Levels of The ACM Computing Classification System (1998)", Association for Computing Machinery.

NOTE: https://www.acm.org/publications/computing-classification-system/1998.

[i.27] Yim, J., Chopra, R., Spitz, T., Winkens, J., Obika, A., Kelly, C., Askham, H., Lukic, M., Huemer, J., Fasler, K. and Moraes, G., 2020: "Predicting conversion to wet age-related macular degeneration using deep learning". Nature Medicine, pp.1-8.

NOTE: https://doi.org/10.1038/s41591-020-0867-7.

[i.28] McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.C., Darzi, A. and Etemadi, M., 2020: "International evaluation of an AI system for breast cancer screening", Nature, 577(7788), pp.89-94.

NOTE: https://doi.org/10.1038/s41586-019-1799-6.

[i.29] Massachusetts Institute of Technology (MIT): "Moral Machine".

NOTE: http://www.moralmachine.net.

[i.30] Organisation for Economic Co-operation and Development (OECD) Council recommendation on Artificial Intelligence.

NOTE: https://www.oecd.org/going-digital/ai/principles/.

[i.31] Chatbot which mimicked the speaking style of characters from a famous television show.

NOTE: https://www.vox.com/2016/4/24/11586346/silicon-valley-hbo-chatbots-for-season-3-premier.

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

full knowledge attack: attack carried out by an attacker who has full knowledge of the system inputs and outputs and its internal design and operations

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

opaque system: system or object which can be viewed solely in terms of its input, output and transfer characteristics without any knowledge of its internal workings

partial knowledge attack: attack carried out by an attacker who has full knowledge of the system inputs and outputs, but only a limited understanding of its internal design and operations

zero knowledge attack: attack carried out by an attacker who has knowledge of the system inputs and outputs, but no knowledge about its internal design or operations

3.2 Symbols

Void.

3.3 Abbreviations

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

ACM	Association for Computing Machinery
AI	Artificial Intelligence
ASIC	Application Specific Integrated Circuit
CCTV	Closed Circuit Television
CNN	Convolutional Neural Network
CVF	Computer Vision Foundation
EPFL	École Polytechnique Fédérale de Lausanne
FPGA	Field Programmable Gate Array
GPU	Graphics Processing Unit
HTML	Hyper Text Markup Language
IEEE	Institute of Electrical and Electronics Engineers
ITU	International Telecommunications Union
OECD	Organisation for Economic Co-operation and Development
RNN	Recurrent Neural Network
TEE	Trusted Execution Environment
UN	United Nations

4 Context

4.1 History

The term 'artificial intelligence' originated at a conference in the 1950s at Dartmouth College in Hanover, New Hampshire, USA. At that time, it was suggested that true artificial intelligence could be created within a generation. By the early 1970s, despite millions of dollars of investment, it became clear that the complexity of creating true artificial intelligence was much greater than anticipated, and investment began to drop off. The years that followed are often referred to as an 'AI winter' which saw little interest or investment in the field, until the early 1980s when another wave of investment kicked off. By the late 1980s, interest had again waned, largely due to the absence of sufficient computing capacity to implement systems, and there followed a second AI winter.

In recent years, interest and investment in AI has once again surfaced, due to the implementation of some practical AI systems enabled by:

- The evolution of advanced techniques in machine learning, neural networks and deep learning.
- The availability of significant data sets to enable robust training.
- Advances in high performance computing enabling rapid training and development.
- Advances in high-performance devices enabling practical implementation.

After the emergence of practical AI systems, suggested theoretical attacks on such systems have become plentiful. However, real-world practical attacks with sufficient motivation and impact are less common.

4.2 AI and machine learning

The field of artificial intelligence is broad, so in order to identify the issues in securing AI, the first step is to define what AI means.

The breadth of the field creates a challenge when trying to create accurate definitions.

EXAMPLE: The Association for Computing Machinery (ACM) Computing Classification System [i.26], Artificial Intelligence is broken down into eleven different categories, each of which has multiple sub-categories.

This represents a complex classification system with a large group of technology areas at varying stages of maturity, some of which have not yet seen real implementations, but does not serve as a useful concise definition. For the purposes of the present document, the following outline definition is used:

- **Artificial intelligence** is the 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.

This definition still represents a broad spectrum of possibilities. However, there are a limited set of technologies which are now becoming realisable, largely driven by the evolution of machine learning and deep learning techniques. Therefore, the present document focusses on the discipline of machine learning and some of its variants, including:

- **Supervised learning** - where all the training data is labelled, and the model can be trained to predict the output based on a new set of inputs.
- **Semi-supervised learning** - where the data set is partially labelled. In this case, even the unlabelled data can be used to improve the quality of the model.
- **Unsupervised learning** - where the data set is unlabelled, and the model looks for structure in the data, including grouping and clustering.
- **Reinforcement 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.

Within each of these machine learning paradigms, there are various model structures that might be used, with one of the most common approaches being the use of deep neural networks, where learning is carried out over a series of hierarchical layers that mimic the behaviour of the human brain.

There are also a number of different training techniques which can be used, including adversarial learning, where the training set contains not only samples which reflect the desired outcomes, but also adversarial samples, which are intended to challenge or disrupt the expected behaviour.

4.3 Data processing chain (machine learning)

4.3.1 Overview

The question of securing AI systems can be simply stated as ensuring the confidentiality, integrity and availability of those systems throughout their lifecycle. The life cycle for machine learning can be considered to have the following stages, as shown in Figure 1.

1) Data acquisition
2) Data curation
3) Model design
4) Software Build
5) Train