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

GROUP REPORT

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

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

6) Test

7) Deployment

8) Updates

Stages 4), 5) and 6) (Build, Train, Test) can together be considered as an iterative implementation cycle.

In the machine learning lifecycle, the training phase can be considered as the most critical, since it is this stage that establishes the baseline behaviour of the system.

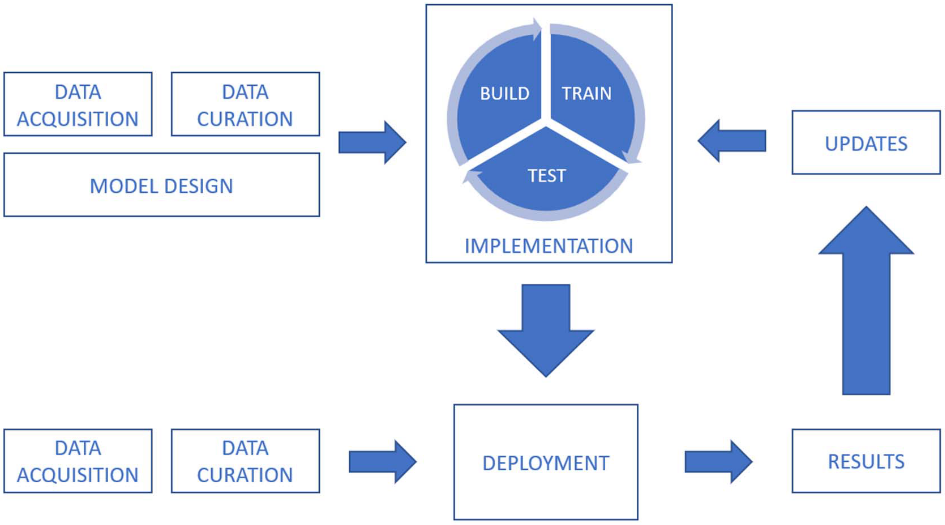

Figure 1: Typical machine learning lifecycle

The level of activity within each phase is dependent on the type of machine learning being used.

EXAMPLE: In unsupervised learning, there is no requirement for data labelling within the data curation stage.

The following clauses address the challenges of ensuring confidentiality, integrity and availability as they apply within those specific stages, with a summary shown in Table 1.

NOTE: The following clauses consider only those challenges which are specific to machine learning systems, and do not consider challenges which are generic across all hardware and software systems.

Table 1: Challenges in confidentiality, integrity and availability in the machine learning lifecycle

Clause	Lifecycle Phase	Issues
4.3.2	Data Acquisition	Integrity
4.3.3	Data Curation	Integrity
4.3.4	Model Design	Generic issues only
4.3.5	Software Build	Generic issues only
4.3.6	Train	Confidentiality, Integrity, Availability
4.3.7	Test	Availability
4.3.8	Deployment	Confidentiality, Integrity, Availability
4.3.9	Upgrades	Integrity, Availability

4.3.2 Data Acquisition

4.3.2.1 Description

In an AI system, data can be obtained from a multitude of sources, including sensors (such as CCTV cameras, mobile phones, medical devices) and digital assets (such as data from trading platforms, document extracts, log files). Data can also be in many different forms (including text, images, video and audio) and can be structured or unstructured. In addition to security challenges related to the data itself, it is important to consider the security of transmission and storage.

4.3.2.2 Integrity challenges

The integrity, or quality, of a data set is a critical success factor for machine learning systems. Data can be poisoned through a deliberate malicious act, i.e. a poisoning attack (described in clause 6.1), but can also become 'poisoned' (or degraded) accidentally (e.g. where insufficient care is taken in collecting data which is consistent and fit for purpose).

4.3.3 Data Curation

4.3.3.1 Description

This phase involves preparing the collected data for use with the intended machine learning approach. This can include integrating data from multiple sources and formats, identifying missing components of the data, removing errors and sources of noise, conversion of data into new formats, labelling the data, data augmentation using real and synthetic data, or scaling the data set using data synthesis approaches.

4.3.3.2 Integrity challenges

When repairing, augmenting or converting data sets, it is important to ensure that the processes do not risk impacting on the quality and integrity of the data. For supervised machine learning systems, it is important that the data labelling is accurate and as complete as possible, and to ensure that the labelling retains its integrity and is not compromised, e.g. through poisoning attacks. It is also important to address the challenge of ensuring the data set is unbiased. Techniques for data augmentation can impact on the integrity of the data.

4.3.4 Model Design

The design of a machine learning system usually contains various artefacts including diagrams, equations, cost functions and optimisation algorithms. For complex models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) there can be multiple artefacts representing each layer of the model.

There are no security challenges which are specific to machine learning systems, but generic challenges should be considered.

4.3.5 Software Build

This refers to the specification, design, and implementation of the software, including the use of commercially available tools and languages.

There are no security challenges which are specific to machine learning systems, but generic challenges should be considered.

4.3.6 Training

4.3.6.1 Description

The training phase of the machine learning process is one of the most critical steps, since it establishes the baseline behaviour of the application. This is the area that is most likely to present unique challenges in relation to security, as learning is at the core of the machine learning process.

The training stage consists of running the model iteratively with a baseline data set for which the desired output is known. With each iteration, the model parameters are adjusted to achieve more accurate performance, and this is repeated until an optimal or acceptable level of accuracy is achieved. It is critical that the training data set is of high quality, as any inaccuracies or inconsistencies can lead to the model behaving incorrectly.

It is possible to use third party components to support the algorithm training phase, and these can sometimes be accessed remotely, e.g. through cloud-based services. In all such cases, the security challenges in clauses 4.3.6.2 to 4.3.6.4 should be considered, together with the generic challenges of using external or cloud-based components and services.

4.3.6.2 Confidentiality challenges

The training dataset confidentiality can be compromised by an attacker with some knowledge of the algorithm implementation (full or partial knowledge attack), or even by a malicious actor with no knowledge of the internal operation of the algorithm (zero knowledge attack) [i.6].

EXAMPLE 1: With zero knowledge of the original training data set, or the model parameters, an attacker creates an augmented training data set with malicious synthetic inputs which are specifically designed to output labels containing information about the original training data. When trained with both the original and augmented data sets, the algorithm can be trained to leak information about the original training data in response to the malicious inputs.

EXAMPLE 2: With some knowledge of the model parameters, an attacker can leverage unused bits within the parameters to leak information.

4.3.6.3 Integrity challenges

An integrity vulnerability can sometimes be introduced through the use of transfer learning, where a pre-trained network can be finely tuned with a few training samples for malicious purposes. Although not intended to be so, meta-learning can be argued to introduce vulnerabilities, in that the system is optimised for transfer learning with only a very small number of samples [i.7]. Furthermore, features generated by a model can be used, without further training, for inference purposes other than that which they were intended.

A backdoor vulnerability is when a special pattern is included during the training phase, and then a trigger is used to generate an output during the inference phase [i.8]. This type of attack can include poisoning during the training phase as a component of its attack, but it is important to emphasise that a backdoor attack requires action in both the training and implementation phases, whereas a poisoning attack requires action only in the training phase. Such a backdoor can also be facilitated through transfer learning.

Another vulnerability is to do with the learning algorithm, which can be modified by an attacker with full or partial knowledge of the algorithm to ensure an incorrect inference for certain samples, whilst maintaining correct performance for the client test set.

4.3.6.4 Availability challenges

In this case, the availability of the machine learning model can be compromised by poisoning attacks on the training data set, which result in the wrong inference result. Normally, this refers to poisoning attacks at the input layer, however, poisoning attacks can also be carried out on the algorithm or its associated parameters.

In addition, in unsupervised learning approaches such as clustering, feature selection is an important step. There are techniques for simultaneous feature selection and clustering which result in feature weightings, or saliency measures. If the attacker has full or partial knowledge of these weightings, they could modify feature values to perform an availability attack in which the salient features are made unavailable by reversing the weights. This attack of denying salient features to the system is known as 'denial of features'

4.3.7 Testing

4.3.7.1 Description

During the testing phase, a portion of the training data set (which has been set aside, and not used during the training phase) is used to validate the performance of the model and its parameters. This includes validating that the model operates correctly from a functional perspective, that the training data set has sufficient coverage of expected inputs, and that the parameters have been correctly configured. Adversarial testing can also be used to test that unexpected or previously unseen inputs do not cause the system to malfunction or become unavailable.

In addition, as with traditional software systems, the code which has been written to implement the model also needs to be tested.

4.3.7.2 Availability challenges

Learnt models can be vulnerable to adversarial samples that result in the functionality not meeting the specification and therefore the required function or service not being available to the user. Therefore, it is important that the test set has sufficient coverage from a testing perspective. Test sets can include adversarial test samples generated by an adversary and those naturally occurring through lack of generalisation. Adversarial testing tries to quickly find testing samples that can cause failure [i.18].

A related area, formal verification of machine learning models, can help in ensuring that the system meets the original specification. However, enumerating all possible outputs for a given set of inputs can often be intractable due to the huge number of choices for the inputs. Efficient approaches to formal verification can be obtained by setting geometric bounds on the output [i.19].

Standardisation of adversarial testing and formal verification algorithms will be important in terms of ensuring the robustness of learnt models.

4.3.8 Deployment and Inference

4.3.8.1 Description

Deployment of machine learning systems has the same challenges as generic systems, including choices about architecture, hardware/software deployment and use of features such as Trusted Execution Environments (TEEs). In addition to these and other similarly generic considerations, it is important to consider any additional performance requirements of a machine learning system. For example, the choice of a TEE can provide a better level of protection for the core system components but may not be able to provide the level of performance provided by generic processors or GPUs.

Hardware deployments are attractive due to high levels of performance and are being explored for both the deployment of machine learning systems and by attackers wishing to exploit vulnerabilities.

As well as risks to the deployed system, there are also certain vulnerabilities that can impact confidentiality by revealing information about the machine learning model, or the data used to train it, as described below.

4.3.8.2 Confidentiality challenges

The main vulnerability in relation to the deployment of machine learning models is their susceptibility to a back-door attack that can compromise the confidentiality of the training set. The nature of the attack depends on whether the attacker has zero, partial or full knowledge of the implementation and operation of the model.

Related to training dataset confidentiality are membership inference and model inversion issues. Membership inference [i.24] means that given a data sample and access to the model, one can distinguish whether this data sample is included in the training dataset. Model inversion [i.25] means that given a prediction result and access to the model, one can find the inference input. Similarly, they are also highly related to the attack type of reverse engineering.

There are also vulnerabilities in relation to the model confidentiality when deployed on untrusted devices.

4.3.8.3 Integrity challenges

The integrity of a machine learning algorithm can be impacted by an attacker who forces the system to behave in an unexpected manner, or to return incorrect results.

This can be carried out directly while the system is in operation by carrying out an input attack as described in clause 6.2.

The integrity of the system can also be impacted through a back-door attack, which has been embedded as part of the training phase, and then triggered by a specific combination of inputs during the inference phase, as described in clause 6.3.

4.3.8.4 Availability challenges

The availability of a system can be affected by various types of attack which cause catastrophic failure, or which cause the system to malfunction in such a way that it becomes unavailable.

Evasion attacks (such as malware obfuscation) can be used to introduce malicious behaviour into the system causing a denial of service.

4.3.9 Upgrades

4.3.9.1 Description

Upgrades to the machine learning model or the machine learning inference software should be treated with the same level of care as a generic change to a deployed system.

In addition, updates to the model parameters should be managed carefully, as any compromise can result in integrity or availability issues.

4.3.9.2 Integrity challenges

Integrity issues can be caused by back door attacks which are deployed during the training phase, and then triggered by an update to system parameters or model.

4.3.9.3 Availability challenges

Updated model parameters can be vulnerable to a poisoning attack as described in clause 6.1.

5 Design challenges and unintentional factors

5.1 Introduction

This clause describes challenges that need to be considered in the design of AI systems, or which can be the result of unintended consequences. They can be considered as security challenges or as features that can be exploited by malicious actors.

5.2 Bias

For machine learning applications, it is advantageous to have a large and well-balanced data set, and to ensure that the decisions made by the machine learning application are not prejudiced or biased in any way. Bias can manifest itself in several different ways.

- **Confirmation bias** occurs when data is selected or manipulated so that it produces outputs aligned to some predetermined assumptions.

- **Selection bias** occurs when data is selected subjectively, resulting in a data set that does not accurately reflect the population.

EXAMPLE 1: When data is gathered using a survey, those people who are willing to participate in the survey are not necessarily an accurate reflection of the entire population.

- **Outliers** are data points which contain extreme values, and therefore can have a disproportionate impact.

EXAMPLE 2: When analysing customer spending habits, the presence of a single customer who spends significantly more than all the others will impact heavily on the average.

- **Underfitting** (where a model is too simplistic) and **overfitting** (where a model is overly complex) can both lead to an inaccurate view of the real data.

It is important to distinguish between systems that display unintended bias, and those whose design displays specific tendencies. Such tendency towards certain behaviours should be regarded as design goals for the system, not as bias.

EXAMPLE 3: In a safety-critical application, such as autonomous vehicles, it can be desirable for the system to exhibit behaviour which errs on the side of safety rather than risk.

EXAMPLE 4: In some systems, greater attention is paid to outliers, as they represent very important instances that have an unusually high impact on behaviour.

It is also important that bias is considered not only during the design and training phases, but also how bias can be introduced after a system has been deployed. In a famous example in 2016, a chatbot was launched, which was intended as an experiment in "conversational understanding", where the chatbot would engage with social networks users through tweets and direct messages. Within a matter of hours, the chatbot was beginning to tweet highly offensive messages. After the chatbot was withdrawn, it was discovered that the chatbot's account had been manipulated to display biased behaviour by internet trolls [i.15].

Bias does not necessarily represent a security issue, but can simply result in the system not meeting its functional requirement.

5.3 Ethics

5.3.1 Introduction

The very concept of artificial intelligence introduces a number of ethical questions, although the perspective on these can be heavily influenced by culture, religion, philosophy and other factors. When it comes to the implementation of specific solutions, ethical questions can be more focused, as in the application of AI to identity and surveillance. There are also sector-specific ethical concerns, such as using AI-based solutions within the healthcare or justice sectors.

5.3.2 Ethics and security challenges

5.3.2.1 Access to data

One of the major ethical concerns of AI systems relates to data privacy, in particular the use which is made of a consumer's data within an AI system. Some AI systems, such as virtual home assistants require the collection, analysis and processing of data not only to make decisions, but as training data to refine and improve the services they offer. While it is clearly necessary and acceptable for a consumer to provide data in order to make specific decisions or provide guidance, it could be considered as unethical to use that data for generic training which is beyond the original purpose. In addition, if the data gathered through such virtual assistants is business-related data, then data leakage can be a commercial risk, by leading to the exposure of intellectual property or business-sensitive information.

It should be noted that there are many AI systems that operate without using personal or sensitive data, in which case concerns about data privacy are less relevant.

Another challenge occurs when the concern over privacy has an impact on the performance or accuracy of a system.

EXAMPLE 1: In the field of medical diagnosis, it is beneficial to have as broad a range of training data as possible. If many individuals choose to withhold their data due to privacy concerns, the effectiveness of the training could be impacted, resulting in a less efficient and effective system.

This applies also to cases where data is not of a personal nature but can be confidential or commercially sensitive.

EXAMPLE 2: A company uses their customer data to generate some insights from a machine learning algorithm, but is unable to use the same data to further train and improve the model.

5.3.2.2 Decision-making

The second problem is related to the ethics and humanity of decision-making in AI systems, which is an extremely complicated problem. On the one hand, it could be considered unethical to use technology for making life-impacting decisions, since an electronic system has no moral or ethical compass to guide its decisions. On the other hand, it could be considered unethical **not** to use the latest technology to support decision-making.

A paper from University of Brighton [i.20] discussed a hypothetical scenario where a car powered by AI knocks down a pedestrian, and explored the legal liabilities that ensue. In March 2018, this scenario became a reality when a self-driving car hit and killed a pedestrian in the city of Tempe, Arizona. This brought into sharp focus not only the legal liabilities, but the potential ethical challenges of the decision-making process itself. In 2016, Massachusetts Institute of Technology (MIT) launched a web site called Moral Machine [i.29] exploring the challenges of allowing intelligent systems to make decisions that are of an ethical nature. The site attempts to explore how humans behave when faced with ethical dilemmas, and to gain a better understanding of how machines ought to behave. The scenarios on which visitors can experiment include choosing the outcome of an accident involving a self-driving car and deciding between outcomes where varying numbers of people are killed or saved.

In the world of healthcare, many companies have invested large amounts of money in developing AI systems for diagnosis of disease. Some of these systems have shown remarkable performance and demonstrate clearly the capability for AI systems to outperform humans. For example, a system using clinical data from an eye hospital in the UK was shown to perform better than five out of six experts [i.27]. Another system in the area of breast cancer screening using data from UK and USA was shown to outperform six expert radiologists [i.28]. However, not every system results in such performance enhancement. There are occasions when such systems perform well during the training and testing phases, but perform less well when faced with new patients and the challenges of diagnosis in real-time and under real-world conditions, where no two patients are alike, and the existence of multiple diseases adds to the complexity of decision-making. One analysis of such a system in Korea in 2018 discovered that the diagnoses of the AI system aligned with that of the medical experts only 50 % of the time [i.21]. From an ethical perspective, it is important to consider the fundamental question of whether a system should be used in a certain scenario, and then there is an ethical imperative to ensure that any such deployed systems are trained using sufficient and accurate training data, and to ensure that they continue to perform in a highly accurate and effective manner.

Professor Ronald C. Arkin, from Georgia Institute of Technology, has published several papers about the use of AI in warfare. In one such paper, he makes the case that automated systems can perform more ethically than humans in some battlefield scenarios [i.22].

5.3.2.3 Obscurity

A 2017 report from Pega Systems titled "What Consumers Really Think About AI: A Global Study" [i.23] revealed that many people had significant reservations about the use of AI systems (only 36 % were comfortable with the idea), but that many people actually used such systems without even realising (up to 84 %). Such reservations can be application-specific, with many people having reservations about the use of AI in life-critical scenarios (such as autonomous driving) but having much less concern about use in lifestyle applications (such as a recommender system for books).

5.3.2.4 Summary

While ethical concerns do not have a direct bearing on the traditional security characteristics of confidentiality, integrity and availability, they can have a significant effect on an individual's perception of whether a system can be trusted. It is therefore essential that AI system designers and implementers consider the ethical challenges and seek to create robust ethical systems that can build trust among users.

5.3.3 Ethics guidelines

While there may not be broadly accepted solutions to the challenges set out in the previous clauses, there are many organisations who have produced ethical guidelines for the implementation of AI systems, including governments, researchers and other agencies at both national and international level. However, it is questionable whether the guidelines themselves have had any significant impact on human decision-making in the field of AI. This question was posed by Thilo Hagendorff from the University of Tübingen, Germany in a February 2020 paper [i.17] where he analysed and compared 22 sets of published guidelines, highlighting some overlaps and omissions.

In May 2019, the Organisation for Economic Co-operation and Development (OECD) adopted a set of principles on AI when they approved the OECD Council recommendation on Artificial Intelligence [i.30]. The principles aimed to set standards for AI that were practical and flexible enough to stand the test of time in a rapidly evolving field, and complemented other OECD standards in areas such as privacy, digital security, risk management and responsible business conduct.

In April 2019, the European Commission High Level Expert Group on Artificial Intelligence published their Ethics Guidelines for Trustworthy AI [i.4]. In their report, they recommended four Ethical Principles that can be considered as the foundations for lawful, ethical and robust AI systems, seven Key Requirements when implementing AI systems, and a series of critical concerns raised by AI.

Since 2017, the United Nations has been holding an annual AI for Good Summit, organised by the International Telecommunication Union (ITU), UN sister agencies, XPRIZE Foundation and ACM. The week-long event brings together business, government and civil society to identify practical applications of AI, and scale those for global impact. The ITU also has a number of related groups exploring AI and machine learning:

- ITU-T Focus Group on "Artificial Intelligence for Health" (FG-AI4H)
- ITU-T Focus Group Machine Learning for Future Networks including 5G (FG-ML5G)
- ITU-T Focus Group on AI for autonomous and assisted driving (FG-AI4AD)
- ITU-T Focus Group on "Environmental Efficiency for Artificial Intelligence and other Emerging Technologies" (FG-AI4EE)

In the UK, in 2019, the Alan Turing Institute published a guide for the responsible design and implementation of AI systems in the public sector [i.3], with the intention of assisting delivery leads in ensuring they develop and deploy AI ethically, safely and responsibly. It is designed to complement and supplement the UK Government's Data Ethics Framework [i.5].

In Germany, in 2019, the Datenethikkommission (data ethics commission) published a report addressing the ethics of AI [i.16]. It contains ethical criteria and policy recommendations for protecting the individual, fostering social coexistence and for securing and promoting wealth in the digital age. The report considers AI as a special case of an algorithmic system and stresses that most of its findings apply to algorithmic systems in general.

5.4 Explainability

In order to trust the behaviour of AI systems, it is important that the decision-making processes are transparent, understandable and explainable. The level of explainability required can often be related to the application in question. For simple recommender systems, where the results are not life-critical, it may be satisfactory to trust the outputs of an AI system without any real understanding of how or why the decision was made. However, for life-critical decisions (such as those in the healthcare or autonomous vehicle domains) it is not only critical that the decisions made by the system are accurate and can be trusted, but that they are also transparent. This requirement for explainability can also occur in heavily regulated industries (such as financial services) where systems need to demonstrate regulatory compliance, but also need to be transparent about how they have achieved that compliance.

Achieving a sufficient level of explainability can be straightforward when using basic machine learning approaches (such as decision trees) but becomes a much more serious challenge when more complex approaches (such as neural networks and deep learning) are adopted.

Explainability is more related to assurance and trust than to security.

5.5 Software and hardware

Traditional software vulnerabilities also exist in AI systems and should be treated in the same way as for any software system.

Hardware implementations are becoming more common, particularly for applications where the training or deployment requires very high-speed processing, and systems might be deployed using GPU, FPGA or ASIC implementations. This clearly introduces an entirely different threat surface, a full analysis of which is outside the scope of the present document.

6 Attack types

6.1 Poisoning

In a poisoning attack, an attacker seeks to compromise the AI model, normally during the training phase, so that the deployed model behaves in a way that the attacker desires. This can be due to the model failing based on certain tasks or inputs, or that the model learns a set of behaviours that are desirable for the attacker, but not intended by the model designer.

Poisoning attacks can typically occur in three ways:

- **Data set poisoning** is often the most direct way to poison a model, since the data set contains all of the knowledge on which the model is based. If an attacker can introduce incorrect, or incorrectly labelled, data into the data set, then the entire learning process can be disrupted. This can be done during the data collection or data curation phases, and can be very hard to detect, since training data sets are typically very large, and often come from multiple distributed sources.

- **Algorithm poisoning** occur when an attacker interferes with the algorithms used for the learning process. For example, federated learning is an approach which aims to protect the privacy of an individual's data. It does this by training individual models on subsets of data, and then combining the learned models together to form the final model. This means the individual data sets remain private but creates an inherent vulnerability. Since any individual data set could be controlled by an attacker, they could manipulate that part of the learning model directly and influence the overall learning of the system.

- **Model poisoning** occurs when the entire deployed model is simply replaced by an alternative model. This type of attack is similar to a traditional cyber-attack where the electronic files comprising the model could be altered or replaced.

6.2 Input attack and evasion

An input attack (also referred to as an evasion attack) occurs when an attacker modifies the input to the AI system to cause the system to malfunction. Such changes or perturbations can be very small, making them very hard, if not impossible, to detect. For example, by changing just a few pixels of an input image, the system might be forced to wrongly identify the image. Another example, presented at the IEEE/CVF Conference on Computer Vision and Pattern Recognition in 2018 [i.10] shows how an adversary can make a small change to a traffic sign, and completely change how a system will interpret it.

Input attacks occur in the deployment phase when systems are already in use and do not require the integrity of the system itself to be compromised at all. The AI system simply behaves as it should, with the output being manipulated due to specific changes in the input.

6.3 Backdoor Attacks

Backdoor attacks refer to attacks where an attacker can:

1) embed special patterns in the model during training phase; and

2) trigger an unexpected output by a designed input (called triggers) during the inference phase.

Backdoor attacks therefore involve both the training and inference phases, whereas poisoning attacks and evasion attacks involve only a single phase (either the training or inference phase). A backdoor attack can use poisoning as part of the attack, but it is not necessary. Other means to conduct backdoor attacks include via transfer learning, where a student model inherits backdoors from the teacher model.

A backdoor attack may attempt to introduce very specific behaviour and outputs from the AI system, or it may be a more general attack, for example an attempt to redirect resources or degrade performance, which could ultimately lead to catastrophic failure.

6.4 Reverse Engineering

Most AI systems are opaque, where the systems accept inputs, and generate outputs without ever revealing the internal logic, algorithms or parameters. In addition, training data sets, which effectively contain all the knowledge of the trained system, are also usually kept confidential. This means it is usually impossible for an outside observer to determine exactly how a system works, or why it produces particular outputs. However, even the most carefully protected systems can be susceptible to reverse engineering.

For example, in a 2016 paper, researchers from Cornell Tech, EPFL, and University of North Carolina showed that it was possible to use a so-called "model extraction attack" to effectively reproduce the functionality of a machine learning system [i.11].

More recently, in 2018, researchers from the Max-Planck Institute for Informatics showed how they could infer information from opaque models by using a sequence of input-output queries [i.12].

7 Misuse of AI

There are two general areas of potential misuse related to AI systems, firstly the misuse of data, and secondly the misuse of the AI algorithms themselves.

With respect to data, where the data gathered for training or input to the system is of a personal or sensitive nature, there are certain rights and protections provided by regulation, whereby such data can only be used for the stated purpose. Misuse of such data could occur unintentionally and without any malicious intent.

EXAMPLE 1: An AI system evolves to produce outputs that are different from the original system.

In such cases, care is essential so that the use of personal or sensitive data continues to conform to the agreed purpose.

A more malicious type of misuse is where an AI algorithm itself is used for a completely different purpose than originally intended.

EXAMPLE 2: An AI system generates authentic human speech. This has many positive applications but can also be used by a malicious actor to attack a voice-based biometric authentication system.

In a 2018 report entitled "The Malicious Use of Artificial Intelligence: Forecasting, Prevention and Mitigation" [i.14], these challenges are set out in some detail by researchers from the Universities of Oxford, Cambridge, Yale, Stanford and others.

8 Real world use cases and attacks

8.1 Overview

This clause describes real-world attacks that have been observed on AI systems. Many theoretical adversarial attacks are of academic interest only and have not been realised in the real world, due to lack of motivation or the lack of reward for the attacker.

8.2 Ad-blocker attacks

Online advertising is a powerful and effective medium, but users can often perceive advertisements as intrusive or malicious. This has led to the growing use of ad-blockers, which can attempt to detect, filter and block ads, often based on the regulatory requirements for ads to be clearly recognisable. This in turn has led to advertisers pursuing creative approaches to avoid detection, for example, by obfuscating the HTML or meta-data associated with the ad.

Increasingly, both participant in the resulting 'arms race' are making use of machine learning approaches, on the one hand to create more effective blocking mechanisms, and on the other hand to find ever more creative ways of avoiding detection.

Perceptual ad-blocking is an approach aimed at identifying ads from their content, rather than from the metadata such as URLs and markup. At the ACM Conference on Computer and Communications Security (CCS) in November 2019, Tramèr et al. demonstrated a security analysis of the perceptual ad-blocking approach, nine general classes of attack against perceptual ad-blocking, and a number of adversarial examples for eight ad classifiers currently available on the market [i.1].

8.3 Malware Obfuscation

The ongoing battle between malware authors and detection tools is another good example of the use of adversarial machine learning, where malware authors use more complex obfuscation techniques to evade detection by machine learning and signature-based tools.

At the 10th ACM Conference on Data and Applications Security and Privacy in early 2020, Millar et al. describe a novel malware detection model for a mobile OS using a deep learning Discriminative Adversarial Network (DAN) [i.2]. This approach is demonstrated to be robust against four real-world obfuscation techniques and demonstrates the potential of the approach to generalise over future obfuscation methods not seen during the training phase. The model is tested against almost 70 000 obfuscated and non-obfuscated malicious and benign software samples.

8.4 Deepfakes

Deepfakes are synthetic media in which a person in an existing image or video is replaced with someone else's likeness. The creators often leverage techniques from machine learning and artificial intelligence to create and manipulate media with highly realistic results. They have been used extensively for celebrity fakes, hoaxes, fake news and financial fraud, and also have the capacity to be used as a tool for ransomware attacks, or as a threat to biometric authentication systems.

8.5 Handwriting reproduction

The subject of handwriting recognition and reproduction presents very significant challenges, largely due to the variability of text, including differences between writing implements (such as fountain pen, pencil or ball pen) and material (paper grade and quality). Variability is also introduced by the writers themselves and can be based on speed of writing and environment.

In 2016, a team from University College London demonstrated a system which used a combination of semi-supervised and unsupervised learning to produce specified output text in the handwriting style of a specific individual [i.9].

Even if such a system is secure, it could be misused by a malicious actor.

EXAMPLE: Enabling an attacker to forge an individual's signature or handwriting style to forge documentation or attack a biometric authentication system.

8.6 Human voice

In September 2016, researchers published a paper describing the use of machine learning for generation of audio [i.13]. This showed:

1) how non-existent but human language-like speech can be produced;

2) how the same approach could be used to improve existing text-to-speech capability; and

3) how the approach could be used to generate realistic passages of music.

Such systems can be misused by malicious actors.

EXAMPLE: A system which reproduces realistic-sounding samples of the voice of a specific individual combined with a text-to-speech capability could be used to attack biometric authentication systems based on voice recognition.

8.7 Fake conversation

In 2016, a California start-up built and released a chatbot which mimicked the speaking style of a number of characters from a famous television show [i.31]. This enabled users to have somewhat realistic, although limited, conversations with the chatbot, whose responses were based on hours of audio recordings of speech for three main characters.

In 2017, the chatbot was evolved further to base its speech patterns on the individual consumer who is using it. Although limited in the type of 'intelligence' displayed by common virtual home assistants, the chatbot learns to use similar language, phrasing and intonation as the individual consumer with which it is communicating.

The ability to produce realistic human speech makes it more difficult for an individual to determine whether they are communicating with another human being or a machine, which can lead to different behaviours, and a different approach to security and privacy. Accurate speech reproduction can also be used to attack voice-based biometric authentication systems.

Annex A: Bibliography

- NIST IR 8269: "A Taxonomy and Terminology of Adversarial Machine Learning".

NOTE : This document sets out information related to Attacks (including Targets, Techniques and Knowledge), Defences and Consequences within the field of Adversarial Machine Learning.

History

Document history		
V1.1.1	December 2020	Publication