ETSI TR 104 222 V1.2.1 (2024-07)

TECHNICAL REPORT

Securing Artificial Intelligence;
Mitigation Strategy Report

Reference

RTR/SAI-009

Keywords

artificial intelligence, security

ETSI

650 Route des Lucioles
F-06921 Sophia Antipolis Cedex - FRANCE

Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16

Siret N° 348 623 562 00017 - APE 7112B
Association à but non lucratif enregistrée à la
Sous-Préfecture de Grasse (06) N° w061004871

Important notice

The present document can be downloaded from:
https://www.etsi.org/standards-search

The present document may be made available in electronic versions and/or in print. The content of any electronic and/or print versions of the present document shall not be modified without the prior written authorization of ETSI. In case of any existing or perceived difference in contents between such versions and/or in print, the prevailing version of an ETSI deliverable is the one made publicly available in PDF format at www.etsi.org/deliver.

Users of the present document should be aware that the document may be subject to revision or change of status. Information on the current status of this and other ETSI documents is available at
https://portal.etsi.org/TB/ETSIDeliverableStatus.aspx

If you find errors in the present document, please send your comment to one of the following services:
https://portal.etsi.org/People/CommiteeSupportStaff.aspx

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

Notice of disclaimer & limitation of liability

Copyright Notification

Contents

Intellectual Property Rights

Essential patents

IPRs essential or potentially essential to normative deliverables may have been declared to ETSI. The declarations pertaining to these essential IPRs, if any, are publicly available for **ETSI members and non-members**, and can be found in ETSI SR 000 314: *"Intellectual Property Rights (IPRs); Essential, or potentially Essential, IPRs notified to ETSI in respect of ETSI standards"*, which is available from the ETSI Secretariat. Latest updates are available on the ETSI Web server (https://ipr.etsi.org/).

Pursuant to the ETSI Directives including the ETSI IPR Policy, no investigation regarding the essentiality of IPRs, including IPR searches, has been carried out by ETSI. No guarantee can be given as to the existence of other IPRs not referenced in ETSI SR 000 314 (or the updates on the ETSI Web server) which are, or may be, or may become, essential to the present document.

Trademarks

The present document may include trademarks and/or tradenames which are asserted and/or registered by their owners. ETSI claims no ownership of these except for any which are indicated as being the property of ETSI, and conveys no right to use or reproduce any trademark and/or tradename. Mention of those trademarks in the present document does not constitute an endorsement by ETSI of products, services or organizations associated with those trademarks.

DECT™, **PLUGTESTS™**, **UMTS™** and the ETSI logo are trademarks of ETSI registered for the benefit of its Members. **3GPP™** and **LTE™** are trademarks of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners. **oneM2M™** logo is a trademark of ETSI registered for the benefit of its Members and of the oneM2M Partners. **GSM**® and the GSM logo are trademarks registered and owned by the GSM Association.

Foreword

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

Modal verbs terminology

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

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

1 Scope

The present document summarizes and analyses existing and potential mitigation against threats for AI-based systems as discussed in ETSI GR SAI 004 [i.1]. The goal is to have a technical survey for mitigating against threats introduced by adopting AI into systems. The technical survey shed light on available methods of securing AI-based systems by mitigating against known or potential security threats. It also addresses security capabilities, challenges, and limitations when adopting mitigation for AI-based systems in certain potential use cases.

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".

[i.2] Doyen Sahoo, Quang Pham, Jing Lu and Steven C. H. Hoi: "Online Deep Learning: Learning Deep Neural Networks on the Fly", International Joint Conferences on Artificial Intelligence Organization, 2018.

[i.3] Battista Biggio and Fabio Roli: "Wild patterns: Ten years after the rise of adversarial machine learning", Pattern Recognition, 2018.

[i.4] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu and Victor C. M. Leung: "A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View". IEEE Access 2018.

[i.5] Nicolas Papernot, Patrick D. McDaniel, Arunesh Sinha and Michael P. Wellman: "SoK: Security and Privacy in Machine Learning". IEEE European Symposium on Security and Privacy (EuroS&P) 2018.

[i.6] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang and Anil K. Jain: "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review". International Journal of Automation and Computing volume 17, pages151-178(2020).

[i.7] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay and Debdeep Mukhopadhyay: "Adversarial Attacks and Defences: A Survey", arXiv preprint arXiv:1810.00069v1.

[i.8] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu and Jinwen He: "Towards Privacy and Security of Deep Learning Systems: A Survey". IEEE Transactions on Software Engineering 2020.

[i.9] NIST IR 8269-(Draft): "A Taxonomy and Terminology of Adversarial Machine Learning".

[i.10] Christian Berghoff1, Matthias Neu1 and Arndt von Twickel: "Vulnerabilities of Connectionist AI Applications: Evaluation and Defence", Frontiers in Big Data volume 3, 2020.

[i.11] Blaine Nelson, Marco Barreno, Fuching Jack Chi, Anthony D. Joseph, Benjamin I. P. Rubinstein, Udam Saini, Charles A. Sutton, J. Doug Tygar and Kai Xia: "Exploiting Machine Learning to Subvert Your Spam Filter", Usenix Workshop on Large-Scale Exploits and Emergent Threats (LEET) 2008.

[i.12] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru and Bo Li: "Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning", IEEE Symposium on Security and Privacy 2018: 19-35.

[i.13] Nathalie Baracaldo, Bryant Chen, Heiko Ludwig and Jaehoon Amir Safavi: "Mitigating Poisoning Attacks on Machine Learning Models: A Data Provenance Based Approach", AISec@CCS 2017: 103-110.

[i.14] Sanghyun Hong, Varun Chandrasekaran, Yigitcan Kaya, Tudor Dumitras and Nicolas Papernot: "On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping", arXiv: 2002.11497v2.

[i.15] Nitika Khurana, Sudip Mittal, Aritran Piplai and Anupam Joshi: "Preventing Poisoning Attacks On AI Based Threat Intelligence Systems", IEEE International Workshop on Machine Learning for Signal Processing (MLSP) 2019: 1-6.

[i.16] Battista Biggio, Igino Corona, Giorgio Fumera, Giorgio Giacinto and Fabio Roli: "Bagging Classifiers for Fighting Poisoning Attacks in Adversarial Classification Tasks", International Workshop on Multiple Classifier Systems (MCS) 2011: 350-359.

[i.17] Yao Cheng, Cheng-Kang Chu, Hsiao-Ying Lin, Marius Lombard-Platet and David Naccache: "Keyed Non-parametric Hypothesis Tests", International Conference on Network and System Security (NSS) 2019: 632-645.

[i.18] Tran, Brandon, Jerry Li and Aleksander Madry: "Spectral signatures in backdoor attacks", In Advances in Neural Information Processing Systems, pp. 8000-8010. 2018.

[i.19] Chen, Bryant, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy and Biplav Srivastava: "Detecting backdoor attacks on deep neural networks by activation clustering", Artificial Intelligence Safety Workshop @ AAAI, 2019.

[i.20] Yuntao Liu, Yang Xie and Ankur Srivastava: "Neural Trojans", 2017 IEEE International Conference on Computer Design (ICCD), Boston, MA, 2017, pp. 45-48, doi: 10.1109/ICCD.2017.16.

[i.21] Bingyin Zhao and Yangjie Lao: "Resilience of Pruned Neural Network against poisoning attack", International Conferecne Conference on Malicious and Unwanted Software (MALWARE) 2018, page 78-83.

[i.22] Liu, Kang, Brendan Dolan-Gavitt and Siddharth Garg: "Fine-pruning: Defending against backdooring attacks on deep neural networks", In International Symposium on Research in Attacks, Intrusions and Defenses, pp. 273-294. Springer, Cham, 2018.

[i.23] Wang, Bolun, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng and Ben Y. Zhao: "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks", In 2019 IEEE Symposium on Security and Privacy (SP), pp. 707-723.

[i.24] Wenbo Guo, Lun Wang, Xinyu Xing, Min Du and Dawn Song. 2019: "Tabor: A highly accurate approach to inspecting and restoring trojan backdoors in AI systems", arXiv preprint arXiv:1908.01763v2 (2019).

[i.25] Yansong Gao, Chang Xu, Derui Wang, Shiping Chen, Damith C.Ranasinghe and Surya Nepal: "STRIP: A Defence Against Trojan Attacks on Deep Neural Networks", 2019 Annual Computer Security Applications Conference (ACSAC '19).

[i.26] Chou Edward, Florian Tramèr, Giancarlo Pellegrino: "sentiNet: Detecting Localized Universal Attack Against Deep Learning Systems", The 3rd Deep Learning and Security Workshop (2020).

[i.27] Sakshi Udeshi, Shanshan Peng, Gerald Woo, Lionell Loh, Louth Rawshan and Sudipta Chattopadhyay: "Model Agnostic Defence against Backdoor Attacks in Machine Learning", arXiv preprint arXiv:1908.02203v2 (2019).

[i.28] Bao Gia Doan, Ehsan Abbasnejad and Damith Ranasinghe: "Februus: Input Purification Defense Against Trojan Attacks on Deep Neural Network Systems", The 36th Annual Computer Security Applications Conference (ACSAC 2020).

[i.29] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh and Dhruv Batra: "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization", International Conference on Computer Vision (ICCV'17).

[i.30] Xiaojun Xu, Qi Wang, Huichen Li, Nikita Borisov, Carl A Gunter and Bo Li: "Detecting AI Trojans Using Meta Neural Analysis", IEEE S&P 2021.

[i.31] Huili Chen, Cheng Fu, Jishen Zhao and Farinaz Koushanfar: "DeepInspect: A Black-box Trojan Detection and Mitigation Framework for Deep Neural Networks", IJCAI 2019: 4658-4664.

[i.32] Soheil Kolouri, Aniruddha Saha, Hamed Pirsiavash and Heiko Hoffmann: "Universal Litmus Patterns: Revealing Backdoor Attacks in CNNs", CVPR 2020.

[i.33] Ambra Demontis, Marco Melis, Maura Pintor, Matthew Jagielski, Battista Biggio, Alina Oprea, Cristina Nita-Rotaru and Fabio Roli: "Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks", USENIX Security Symposium 2019.

[i.34] Yinpeng Dong, Tianyu Pang, Hang Su and Jun Zhu: "Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks", CVPR 2019.

[i.35] Carlini and Wagner: "Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods", the 10th ACM Workshop on Artificial Intelligence and Security 2017.

[i.36] Anish Athalye, Nicholas Carlini, David Wagner: "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples", ICML 2018.

[i.37] Florian Tramer, Nicholas Carlini, Wieland Brendel and Aleksander Madry: "On Adaptive Attacks to Adversarial Example Defenses", NIPS 2020.

[i.38] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, Jörn-Henrik Jacobsen: "Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations", ICML 2020.

[i.39] Gintare Karolina Dziugaite, Zoubin Ghahramani and Daniel M. Roy: "A study of the effect of JPG compression on adversarial images", International Society for Bayesian Analysis (ISBA 2016) World Meeting.

[i.40] Uri Shaham, James Garritano, Yutaro Yamada, Ethan Weinberger, Alex Cloninger, Xiuyuan Cheng, Kelly Stanton and Yuval Kluger: "Defending against Adversarial Images using Basis Functions Transformation", ArXiv 2018.

[i.41] Richard Shin and Dawn Song: "JPEG-resistant adversarial images", in Proc. Mach. Learn. Comput. Secur. Workshop, 2017.

[i.42] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E. Kounavis and Duen Horng Chau: "Shield: Fast, Practical Defense and Vaccination for Deep Learning using JPEG Compression", KDD 2018.

[i.43] Chuan Guo, Mayank Rana, Moustapha Cissé and Laurens van der Maaten: "Countering Adversarial Images using Input Transformations", ICLR 2018.

[i.44] Hossein Hosseini, Yize Chen, Sreeram Kannan, Baosen Zhang and Radha Poovendran: "Blocking Transferability of Adversarial Examples in Black-Box Learning Systems", ArXiv 2017.

[i.45] Ian J. Goodfellow, Jonathon Shlens and Christian Szegedy: "Explaining and Harnessing Adversarial Examples", ICLR 2015.

[i.46] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu: "Towards Deep Learning Models Resistant to Adversarial Attacks", ICLR 2018.

[i.47] Madry et al. (2019). RobustML.

[i.48] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin and Nicolas Usunier: "Parseval Networks: Improving robustness to adversarial examples", ICML 2017.

[i.49] Ross and Doshi-Velez: "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients", AAAI 2018.

[i.50] Cem Anil, James Lucas and Roger Grosse: "Sorting out Lipschitz function approximation", ICML 2019.

[i.51] Jeremy Cohen, Elan Rosenfeld and Zico Kolter: "Certified Adversarial Robustness via Randomized Smoothing", ICML 2019.

[i.52] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha and Ananthram Swami: "Distillation as a defense to adversarial perturbations against deep neural networks", 2016 IEEE Symposium S&P.

[i.53] Nicolas Carlini and David Wagner: "Towards evaluating the robustness of neural networks", 2017 IEEE symposium on Security and Privacy (S&P).

[i.54] Akhilan Boopathy, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu and Luca Daniel: "CNN-Cert: An Efficient Framework for Certifying Robustness of Convolutional Neural Networks", AAAI 2019.

[i.55] Ching-Yun Ko, Zhaoyang Lyu, Tsui-Wei Weng, Luca Daniel, Ngai Wong and Dahua Lin: "POPQORN: Quantifying Robustness of Recurrent Neural Networks", ICML 2019.

[i.56] Aleksandar Bojchevski and Stephan Günnemann:"Certifiable Robustness to Graph Perturbations", NIPS 2019.

[i.57] G. Katz, C. Barrett, D. L. Dill, K. Julian and M. J. Kochenderfer: "Reluplex: An efficient SMT solver for verifying deep neural networks", in International Conference on Computer Aided Verification. Springer, 2017, pp. 97-117.

[i.58] X. Huang, M. Kwiatkowska, S. Wang and M. Wu: "Safety verification of deep neural networks", in International Conference on Computer Aided Verification. Springer, 2017, pp. 3-29.

[i.59] O. Bastani, Y. Ioannou, L. Lampropoulos, D. Vytiniotis, A. Nori and A. Criminisi: "Measuring neural net robustness with constraints", in Advances in neural information processing systems, 2016, pp. 2613- 2621.

[i.60] V. Tjeng, K. Y. Xiaoand R. Tedrake: "Evaluating robustness of neural networks with mixed integer programming", in International Conference on Learning Representations (ICLR), 2019.

[i.61] S. Wang, K. Pei, J. Whitehouse, J. Yang and S. Jana: "Efficient formal safety analysis of neural networks", in Advances in Neural Information Processing Systems, 2018, pp. 6367-6377.

[i.62] Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri and Martin Vechev: "AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation", IEEE S&P 2018.

[i.63] G. Singh, T. Gehr, M. Püschel and M. Vechev: "An Abstract Domain for Certifying Neural Networks", Proceedings of the ACM on Programming Languages, vol. 3, no. POPL, p. 41, 2019.

[i.64] Yizhak Yisrael Elboher, Justin Gottschlich and Guy Katz: "An Abstraction-Based Framework for Neural Network Verification", International Conference on Computer Aided Verification 2020.

[i.65] Guy Katz, Derek A. Huang Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, David L. Dill, Mykel J. Kochenderfer and Clark Barrett: "The Marabou Framework for Verification and Analysis of Deep Neural Networks", International Conference on Computer Aided Verification 2019.

[i.66] ERAN: Neural Network Verification Framework.

[i.67] Lily Weng, Pin-Yu Chen, Lam Nguyen, Mark Squillante, Akhilan Boopathy, Ivan Oseledets and Luca Daniel: "PROVEN: Verifying Robustness of Neural Networks with a Probabilistic Approach", ICML 2019.

[i.68] Linyi Li, Xiangyu Qi, Tao Xie, and Bo Li: "SoK: Certified Robustness for Deep Neural Networks", ArXiv, 2020.

[i.69] Shixin Tian, Guolei Yang and Ying Cai: "Detecting Adversarial Examples Through Image Transformation", AAAI 2018.

[i.70] Weilin Xu, David Evans and Yanjun Qi: "Feature squeezing: Detecting adversarial examples in Deep Neural Networks", Network and Distributed Systems Security Symposium 2018.

[i.71] Bo Huang, Yi Wang and Wei Wang: "Model-Agnostic Adversarial Detection by Random Perturbations", International Joint Conference on Artificial Intelligence (IJCAI-19).

[i.72] Xin Li and Fuxin Li: "Adversarial Examples Detection in Deep Networks with Convolutional Filter Statistics", ICCV 2017.

[i.73] Kevin Roth, Yannic Kilcher and Thomas Hofmann: "The Odds are Odd: A Statistical Test for Detecting Adversarial Example", ICML 2019.

[i.74] Dongyu Meng and Hao Chen: "MagNet: A Two Pronged Defense against adversarial examples", ACM Conference on Computer and Communications Security (CCS) 2017.

[i.75] Nicholas Carlini, David Wagner: "MagNet and "Efficient Defenses Against Adversarial Attacks" are Not Robust to Adversarial Examples", arXiv 2017.

[i.76] Faiq Khalid, Hassan Ali, Hammad Tariq, Muhammad Abdullah Hanif, Semeen Rehman, Rehan Ahmed and Muhammad Shafique: "QuSecNets: Quantization-based Defense Mechanism for Securing Deep Neural Network against Adversarial Attacks", IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS) 2019.

[i.77] Sanjay Kariyappa and Moinuddin K. Qureshi: "Improving adversarial robustness of ensembles with diversity training", ArXiv 2019.

[i.78] Thilo Strauss, Markus Hanselmann, Andrej Junginger, Holger Ulmer: "Ensemble Methods as a Defense to Adversarial Perturbations Against Deep Neural Networks", Canadian Conference on Artificial Intelligence 2020.

[i.79] Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph. Stoecklin, Heqing Huang and Ian Molloy: "Protecting Intellectual Property of Deep Neural Networks with Watermarking", ASIACCS'18.

[i.80] Wang, Tianhao and Florian Kerschbaum: "Attacks on digital watermarks for deep neural networks", in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2622-2626. IEEE, 2019.

[i.81] Takemura, Tatsuya, Naoto Yanai and Toru Fujiwara: "Model Extraction Attacks against Recurrent Neural Networks", arXiv preprint arXiv:2002.00123 (2020).

[i.82] Manish Kesarwani, Bhaskar Mukhoty, Vijay Arya and Sameep Mehta: "Model Extraction Warning in MLaaS Paradigm", ACSAC 2018.

[i.83] Huadi Zheng, Qingqing Ye, Haibo Hu, Chengfang Fang and Jie Shi: "BDPL: A Boundary Differentially Private Layer Against Machine Learning Model Extraction Attacks", ESORICS 2019.

[i.84] Mika Juuti, Sebastian Szyller, Samuel Marchal and N. Asokan: "PRADA: Protecting Against DNN Model Stealing Attacks", 2019 IEEE European Symposium on Security and Privacy (EuroS&P).

[i.85] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter and Thomas Ristenpart: "Stealing Machine Learning Models via Prediction APIs", Usenix Security 2016.

[i.86] Liu, Yuntao, Dana Dachman-Soled and Ankur Srivastava: "Mitigating reverse engineering attacks on deep neural networks", in 2019 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), pp. 657-662.

[i.87] Lukas, Nils, Yuxuan Zhang and Florian Kerschbaum: "Deep Neural Network Fingerprinting by Conferrable Adversarial Examples", arXiv preprint arXiv:1912.00888v3, (2020).

[i.88] Cao, Xiaoyu, Jinyuan Jia and Neil Zhenqiang Gong: "IPGuard: Protecting the Intellectual Property of Deep Neural Networks via Fingerprinting the Classification Boundary", ACM ASIA Conference on Computer and Communications Security (ASIACCS), 2021.

[i.89] Nicolas Papernot, Martin Abadi, Ulfar Erlingsson, Ian Goodfellow, Kunal Talwar: "Semi-supervised knowledge transfer for deep learning from private training data", ICLR 2017.

[i.90] Max Friedrich, Arne Köhn, Gregor Wiedemann and Chris Biemann: "Adversarial learning of privacy preserving test representations for de-identification of medical records", Proceeding of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[i.91] Taihong Xiao, Yi-Hsuan Tsai, Kihyuk Sohn, Manmohan Chandraker and Ming-Hsuan Yang: "Adversarial Learning of Privacy-Preserving and Task-Oriented Representations", AAAI 2020.

[i.92] Cynthia Dwork and Aaron Roth: "The algorithmic foundations of differential privacy", Foundations and Trends in Theoretical Computer Science, 2014.

[i.93] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar and Li Zhang: "Deep learning with differential privacy", Proceedings of the 2016 ACM CCS, 2016.

[i.94] Milad Nasr, Reza Shokri and Amir Houmansadr: "Machine learning with membership privacy using adversarial regularization", Proceedings of the 2018 ACM CCS, 2018.

[i.95] Jinyuan Jia, Ahmed Salem, Michael Backes, Yang Zhang and Neil Zhenqiang Gong: "MemGuard: Defending against Black-Box Membership Inference Attacks via Adversarial Examples", ACM CCS 2019.

[i.96] Ziqi Yang, Bin Shao, Bohan Xuan, Ee-Chien Chang and Fan Zhang: "Defending Model Inversion and Membership Inference Attacks via Prediction Purification", ArXiv, arXiv:2005.03915v2, 2020.

3 Definition of terms, symbols and abbreviations

3.1 Terms

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

adversarial examples: carefully crafted samples which mislead a model to give an incorrect prediction

conferrable adversarial examples: subclass of transferable adversarial examples that exclusively transfer with a target label from a source model to its surrogates

distributional shift: distribution of input data changes over time

inference attack: attacks launched from deployment stage

model-agnostic mitigation: mitigations which do not modify the addressed machine learning model

model enhancement mitigation: mitigations which modify the addressed machine learning model

training attack: attacks launched from development stage

transferable adversarial examples: adversarial examples which are crafted for one model but also fool a different model with a high probability