AI Security and Safety: The PRALab Research Experience

Ambra Demontis[1,*], Maura Pintor[1], Luca Demetrio[2], Angelo Sotgiu[3], Daniele Angioni[1], Giorgio Piras[1], Srishti Gupta[1], Battista Biggio[1] and Fabio Roli[2]

[1] *Pattern Recognition and Applications Laboratory (PRALab), Department of Electrical and Electronic Engineering, University of Cagliari, Italy*
[2] *Department of Informatics, Bioengineering, Robotics, and Systems Engineering, University of Genova*
[3] *Consorzio Interuniversitario Nazionale per l'Informatica (CINI)*

Abstract

We present here the main research topics and activities on security, safety, and robustness of machine learning models developed at the Pattern Recognition and Applications (PRA) Laboratory of the University of Cagliari. We have provided pioneering contributions to this research area, being the first to demonstrate gradient-based attacks to craft adversarial examples and training data poisoning attacks. The findings of our research have significantly contributed not only to identifying and characterizing vulnerabilities of such models in the context of real-world applications but also to the development of more trustworthy artificial intelligence and machine learning models. We are part of the ELSA network of excellence for the development of safe and secure AI-based technologies, funded by the European Union.

Keywords
Artificial Intelligence, Security, Safety, Adversarial Machine Learning

1. Research Group

The Pattern Recognition and Applications (PRA) Laboratory was founded in 1996. The PRALab has been active for more than 25 years at the University of Cagliari. Its mission is to address fundamental issues for the development of future pattern recognition systems in the context of real applications, focused on creating secure systems for security applications, as reflected by our motto:

> *there is nothing more practical than a good theory*, by Kurt Lewin.

Our activities can be categorized into four highly-interdependent lines: (i) development of theories to solve problems of fundamental research, including multiple classifier systems (our original expertise) and adversarial machine learning; (ii) application of these theories to solve practical problems, in the research domains of computer vision for video surveillance and ambient intelligence, computer security, biometrics, document and multimedia categorization; (iii) testing and validation of the proposed solutions on real-world data (in-vivo experiments); and (iv) development of prototypes and demonstrators, through which the results of basic research are translated into functional products.

In 2015, some members of the PRA Lab decided to found the company Pluribus One (https://www.pluribus-one.it/), as a spinoff of the research laboratory. Pluribus One is now a well-developed, research-intensive company that designs innovative products and services based on AI technologies for data-driven business and cybersecurity applications, including protection of web services and endpoint devices.

The PRALab team working on AI/ML security, safety, and robustness consists of nine people, including the lab director (Prof. Fabio Roli), an associate professor (Prof. Battista Biggio), three assistant professors (Dr. Ambra Demontis, Dr. Luca Demetrio and Dr. Maura Pintor), a collaborator (Dr. Angelo Sotgiu), and three more Ph.D. students. Our team has provided pioneering contributions in the area of AI/ML security, being the first to demonstrate gradient-based evasion [1] (also known as *adversarial examples*) and poisoning attacks [2], and how to mitigate them, playing a leading role in the establishment and advancement of this research field [3]. In particular, the work in [2] has received the prestigious 2022 ICML Test of Time Award, recognizing its long-lasting impact since 2012, while the work in [3] has received the Best Paper Award and Pattern Recognition Medal from the journal Pattern Recognition.

2. Research Topics

Since the last decade, the usage of artificial intelligence has grown rapidly. Today, it is used by citizens through vocal assistants, autonomous driving cars, and other technologies that are becoming part of our life, but also by organizations to increase their sales and improve their performance, and by governments; for example, for border monitoring. The increasing usage of artificial intelligence raises concerns about its possible impact on society. This concern is shared by the European Union, which recently wrote the EU Artificial Intelligence Act to reg-

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

*Corresponding author.

ambra.demontis@unica.it (A. Demontis)

ulate the usage of AI, ensuring it will not violate any fundamental human right. This regulation subdivides the AI-based approaches into categories depending on the harm they may cause to human rights and regulates each category with a series of requirements to which the system should be compliant. The main requirements for AI used in high-risk applications, such as biometric identification and law enforcement, regard these three pillars:

- Ethics. Individuals with similar characteristics should receive the same response from the system regardless of their gender, ethnicity, and other characteristics that, for ethical reasons, should not affect it;
- Interpretability. The system should provide the user with information about the process used to provide the output;
- Robustness. Attackers should not be able to alter the integrity, availability, and privacy of the AI system, the data used to train it, and the system's outputs.

Our research focuses mostly on the last two pillars. We have demonstrated that robust systems are uttermost important as the security of supervised [3], unsupervised [3], and reinforcement learning [4] systems can be severely affected by well-crafted attacks. Furthermore, AI may be also used *offensively* to perpetrate scams and cyber-crimes [5]. Studying the vulnerability of AI to attacks and forecasting its possible misuse is important to raise awareness about possible related threats but is also essential to understand the underlying reasons behind the vulnerabilities of AI and create more robust systems. In the following, we will present our recent research regarding the robustness of AI, related to threats that can be staged against AI models either at training time or at test time.

2.1. AI Robustness to Test-time Threats

At test time, attackers can threaten AI by manipulating the samples the system will receive as input. In this way, they can force the system to misclassify a sample. For example, they can add a sticker to a street signal representing a stop to have it misclassified as a speed limit. This attack is called evasion, and the input samples used to perpetrate it are called adversarial examples. Our team has been the first to devise gradient-based evasion attacks and show that some popular classifications (like Support Vector Machines and Neural Networks [1]) and feature selection algorithms are vulnerable to this threat [6]. There are different challenges related which hinder the applicability of evasion attacks to developed systems, which also depends on the considered technology and application. For example, it is challenging to construct adversarial examples against models that are not differentiable [7], and it is even more challenging to construct adversarial examples to evade malware detectors. This is because the attacker might compromise their malicious functionalities by modifying the malware. Our researchers developed evasion algorithms able to construct malware that evade the target system while preserving all their functionalities [8, 9, 10, 11]. Another challenge is that attackers often do not know all the details regarding the target systems. Our researchers have shown that effective attacks can nevertheless be developed in this challenging scenario [12, 10, 11]. Notably, these attacks have also been effective on anti-virus solutions hosted at VirusTotal. Understanding to which extent attacks can be efficient and effective is really important to correctly assess the security of machine learning algorithms, avoiding overestimating their robustness. Different methods that should be able to formally verify these technologies' robustness have been proposed; however, they can be applied only to a limited set of AI/ML technologies. Therefore, in most cases, these technologies are evaluated empirically, simulating attacks and evaluating the security of the systems against them [3]. Our team has done different works to help perform empirical evaluations of machine learning algorithms' security. We have provided methodologies and debugging tools that can be used to improve current approaches for empirical security evaluations, especially the one used for high-risk applications, making them more reliable (e.g., by properly tuning the attack hyperparameters and identifying the presence of gradient obfuscation hindering the attack optimization) [13, 14]. Moreover, we have proposed efficient attacks [15], and created and open-sourced a dataset of adversarial patches [16] that can be used to quickly benchmark machine learning models for image classification. Whereas most of our work and the literature focus on evasion attacks, test time attacks can also have different goals. For example, attackers may want to repurpose a model trained a task to solve a different task. This attack is called reprogramming. In one of our recent works, we have explained the main factors that influence the effectiveness of this attack [17].

In our research, we have proposed different strategies to counter evasion attacks. The first consists of increasing the margin in input space, which can be done with different techniques, including adversarial training [18, 19], and regularization [20, 8, 12]. The second consists of detecting the input samples modified by the attacker [21, 22, 23, 19, 24]. We have recently proposed also a defense to counter reprogramming attacks by analyzing the sequence of queries made to the classifier by the same user.

2.2. AI Robustness to Training-time Threats

At training time, attackers can threaten AI by manipulating the samples the system receives to learn to accomplish the task for which it is developed. In this way, they can make the system unable to learn the task correctly [2, 25], thus committing errors at test time. This attack is called training data poisoning. We have been the first to propose a gradient-based poisoning attack showing that Support Vector Machines [2] are highly affected by this threat. This work received the prestigious 2022 ICML Test of Time Award. Our team has also been the first to show that neural networks [26], feature selection methods [27], and clustering algorithms [28] can also be compromised by this attack. The main challenge regarding poisoning attacks is that generating the optimal attack samples requires solving a computationally costly problem [29]. Nevertheless, we have shown that it is possible to find effective, approximate solutions in a fast manner [26, 30]. As the literature about poisoning attacks is rapidly increasing, we have also proposed a survey [29] that systematizes more than 100 papers published in the field in the last 15 years, shedding light on the current limitations and discussing future open research questions.

3. Projects

Our research activities are carried out in the framework of regional, national, and European projects funded by public as well as private initiatives. We had more than twenty-five projects founded between 2012 and 2020. The full list is available at http://pralab.diee.unica.it/en/Projects. Seven of them were funded by the European Commission, and two of them were coordinated by the PRALab. Overall, we received 3 million euros of funding, with half provided by the European Commission. The annual turnover of the laboratory is around four hundred thousand euros.

We have different ongoing projects on AI security:

1. 2023-2026 - The recently-approved Sec4AI4Sec project aims to devise testing and protection methods for AI-enabled components in software security assets. The project will start in the last quarter of 2023.
2. 2022-2025 - ELSA: "European Lighthouse on Secure and Safe AI," funded by the European Union with 7M euros. This project aims to create a European network of excellence for the development of secure and safe AI.
3. 2020-2023 - FFG COMET Module S3AI: "Security and Safety for Shared Artificial Intelligence," funded by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by the Austrian Research Promotion Agency FFG. This project aims to provide the foundations required to build secure, safe, and shared AI systems.
4. 2020-2023 - PRIN 2017 RexLearn: "Reliable and Explainable Machine Learning," funded by the Italian Ministry of Education, University and Research (grant no.2017TWNMH2). This project aims to develop novel learning paradigms, able to take reliable and explainable decisions, and to assess and mitigate the security risks associated with potential misuses of machine learning.

Some other relevant projects are listed in the following:

- 2017-2019 - Research and Innovation Action LETS-CROWD: “Law Enforcement agencies human factor methods and Toolkit for the Security and protection of CROWDs in mass gatherings”. Call: H2020 - SEC-07-FCT-2016-2017. Grant Agreement H2020/N.740466.
- 2015-2018 – Innovation Action DOGANA: “aDvanced sOcial enGineering And vulNerability Assessment Framework”. Call: H2020 – DS 2014-1. Grant Agreement H2020/N.653618.
- 2014-2016 – CSA CyberROAD: “Development of the Cybercrime and Cyberterrorism Research Roadmap”. Call: FP7 – SEC 2013.2.5-1. Grant Agreement FP7-SEC-2013/N.607642.
- 2014-2016 - ILLBuster, “Buster of ILLegal Contents spread by malicious computer networks”. DGHOME - ISEC, Prevention of and Fight Against Crime. Grant Agreement: HOME/2012/ISEC/AG/4000004360.

4. Developed Tools

As explained in the previous section, correctly evaluating the robustness of AI/ML technologies might be challenging. Our researchers have developed different tools that help to perform security evaluation[1]. These tools include SecML [31], a Python library that allows assessing the security evaluation of AI/ML technologies against evasion and poisoning attacks, and an extension of this library, called SecML Malware [32] ad-hoc for Windows malware. For each of them, they have released a tool that allows running security the evaluations through a graphical interface: PandaVision, and ToucanStrike. Furthermore, our researchers have released a tool that allows evaluating is an attack is or not effective in the considered scenario[2].

[1] https://github.com/pralab
[2] https://github.com/pralab/IndicatorsOfAttackFailure

5. Challenges and Perspectives

While research is quickly progressing in AI/ML Security, companies are working on automating the development and operations of ML models (MLOps), without focusing too much on ML security-related issues. In this respect, a relevant challenge for the future will be to extend the current MLOps paradigm to also encompass ML security (towards implementing what we refer to as ML*Sec*Ops). To this end, we plan to incorporate research on security testing, protection and monitoring of AI/ML models into the MLOps development cycle. In particular, we plan to extend our research towards: (i) developing and improving attacks (including evasion, poisoning and privacy threats) for making security testing and validation of AI/ML models more efficient and available for a wider set of application domains; (ii) designing improved defenses with robustness guarantees to protect AI/ML models not only against such attacks but also to enable reliable classification when out-of-distribution data is provided as input; and (iii) designing methods that allow for constantly monitoring if a deployed model is attacked during operation, enabling prompt reaction when needed. We firmly believe that integrating these dimensions into an MLSecOps cycle will definitely help software engineers and developers to seamlessly deploy and maintain more secure, reliable, and trustworthy AI/ML models in practice.

References

[1] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, F. Roli, Evasion attacks against machine learning at test time, in: ECML PKDD, Part III, volume 8190 of *LNCS*, 2013, pp. 387–402.

[2] B. Biggio, B. Nelson, P. Laskov, Poisoning attacks against support vector machines, in: 29th ICML, 2012, pp. 1807–1814.

[3] B. Biggio, F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, Patt. Rec. 84 (2018) 317–331.

[4] A. Demontis, M. Pintor, L. Demetrio, K. Grosse, H.-Y. Lin, C. Fang, B. Biggio, F. Roli, A Survey on Reinforcement Learning Security with Application to Autonomous Driving, 2022. URL: http://arxiv.org/abs/2212.06123. doi:10.48550/arXiv.2212.06123, arXiv:2212.06123 [cs].

[5] Y. Mirsky, A. Demontis, J. Kotak, R. Shankar, D. Gelei, L. Yang, X. Zhang, M. Pintor, W. Lee, Y. Elovici, B. Biggio, The Threat of Offensive AI to Organizations, Computers & Security 124 (2023) 103006. URL: https://www.sciencedirect.com/science/article/pii/S0167404822003984. doi:10.1016/j.cose.2022.103006.

[6] F. Zhang, P. Chan, B. Biggio, D. Yeung, F. Roli, Adversarial feature selection against evasion attacks, IEEE Trans. on Cybernetics 46 (2016) 766–777.

[7] B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, F. Roli, Adversarial malware binaries: Evading deep learning for malware detection in executables, in: 2018 26th EUSIPCO, IEEE, IEEE, Rome, 2018, pp. 533–537. doi:10.23919/EUSIPCO.2018.8553214.

[8] A. Demontis, M. Melis, B. Biggio, D. Maiorca, D. Arp, K. Rieck, I. Corona, G. Giacinto, F. Roli, Yes, machine learning can be more secure! a case study on android malware detection, IEEE Trans. Dependable and Secure Computing (2019).

[9] B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, F. Roli, Adversarial Malware Binaries: Evading Deep Learning for Malware Detection in Executables, ArXiv (2018). arXiv:1803.04173.

[10] L. Demetrio, B. Biggio, G. Lagorio, F. Roli, A. Armando, Functionality-preserving black-box optimization of adversarial windows malware, IEEE Transactions on Information Forensics and Security 16 (2021) 3469–3478.

[11] L. Demetrio, S. E. Coull, B. Biggio, G. Lagorio, A. Armando, F. Roli, Adversarial EXEmples: A survey and experimental evaluation of practical attacks on machine learning for windows malware detection, ACM Trans. Priv. Secur. 24 (2021).

[12] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, F. Roli, Why do adversarial attacks transfer? Explaining transferability of evasion and poisoning attacks, in: USENIX Security, USENIX Association, 2019.

[13] M. Pintor, L. Demetrio, A. Sotgiu, A. Demontis, N. Carlini, B. Biggio, F. Roli, Indicators of Attack Failure: Debugging and Improving Optimization of Adversarial Examples, in: A. H. Oh, A. Agarwal, D. Belgrave, K. Cho (Eds.), Advances in Neural Information Processing Systems, 2022. URL: https://openreview.net/forum?id=Y1sWzKW0k4L.

[14] M. Pintor, L. Demetrio, G. Manca, B. Biggio, F. Roli, Slope: A First-order Approach for Measuring Gradient Obfuscation, in: Proc. of the ESANN, ESANN 2021, 2021.

[15] M. Pintor, F. Roli, W. Brendel, B. Biggio, Fast Minimum-norm Adversarial Attacks through Adaptive Norm Constraints, in: A. Beygelzimer, Y. Dauphin, P. Liang, J. W. Vaughan (Eds.), Advances in Neural Information Processing Systems, 2021.

[16] M. Pintor, D. Angioni, A. Sotgiu, L. Demetrio, A. Demontis, B. Biggio, F. Roli, ImageNet-Patch: A dataset for benchmarking machine learning robustness against adversarial patches, Pattern Recognition 134 (2023) 109064. URL: https://www.sciencedirect.com/science/article/pii/S0031320322005441. doi:10.1016/j.patcog.2022.109064.

[17] Y. Zheng, X. Feng, Z. Xia, X. Jiang, A. Demontis, M. Pintor, B. Biggio, F. Roli, Why Adversarial Reprogramming Works, When It Fails, and How to Tell the Difference, Information Sciences (2023). URL: https://www.sciencedirect.com/science/article/pii/S0020025523002803. doi:10.1016/j.ins.2023.02.086.

[18] S. Rota Bulò, B. Biggio, I. Pillai, M. Pelillo, F. Roli, Randomized prediction games for adversarial machine learning, IEEE Trans. on Neural Networks and Learning Systems 28 (2017) 2466–2478.

[19] D. Maiorca, A. Demontis, B. Biggio, F. Roli, G. Giacinto, Adversarial Detection of Flash Malware: Limitations and Open Issues, Computers & Security 96 (2020). URL: https://www.sciencedirect.com/science/article/pii/S0167404820301760?dgcid=rss_sd_all. doi:https://doi.org/10.1016/j.cose.2020.101901.

[20] A. Demontis, P. Russu, B. Biggio, G. Fumera, F. Roli, On security and sparsity of linear classifiers for adversarial settings, in: Joint IAPR Int'l Works. on Struct., Synt., and Stat. Patt. Rec., volume 10029, Cham, 2016, pp. 322–332.

[21] M. Melis, A. Demontis, B. Biggio, G. Brown, G. Fumera, F. Roli, Is deep learning safe for robot vision? adversarial examples against the icub humanoid, in: 2017 VIPAR, 2017, pp. 751–759. doi:10.1109/ICCVW.2017.94.

[22] A. Sotgiu, A. Demontis, M. Melis, B. Biggio, G. Fumera, X. Feng, F. Roli, Deep neural rejection against adversarial examples, EURASIP Journal on Information Security 2020 (2020) 5. URL: https://doi.org/10.1186/s13635-020-00105-y. doi:10.1186/s13635-020-00105-y.

[23] F. Crecchi, M. Melis, A. Sotgiu, D. Bacciu, B. Biggio, FADER: Fast adversarial example rejection, Neurocomputing 470 (2022) 257–268. URL: https://www.sciencedirect.com/science/article/pii/S0925231221015708. doi:10.1016/j.neucom.2021.10.082.

[24] S. Melacci, G. Ciravegna, A. Sotgiu, A. Demontis, B. Biggio, M. Gori, F. Roli, Domain Knowledge Alleviates Adversarial Attacks in Multi-Label Classifiers, IEEE Trans. on Patt. Analysis and Machine Intelligence (2021) 1–1. doi:10.1109/TPAMI.2021.3137564.

[25] A. E. Cinà, K. Grosse, S. Vascon, A. Demontis, B. Biggio, F. Roli, M. Pelillo, Backdoor learning curves: Explaining backdoor poisoning beyond influence functions, 2021. arXiv:2106.07214.

[26] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, F. Roli, Towards poisoning of deep learning algorithms with back-gradient optimization, in: Proc. of the 10th ACM Works. AISec@CCS 2017, 2017, pp. 27–38. URL: https://doi.org/10.1145/3128572.3140451. doi:10.1145/3128572.3140451.

[27] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, F. Roli, Is feature selection secure against training data poisoning?, in: ICML, 2015, pp. 1689–1698.

[28] B. Biggio, I. Pillai, S. R. Bulò, D. Ariu, M. Pelillo, F. Roli, Is data clustering in adversarial settings secure?, in: Proc. of the 2013 AISec, AISec '13, New York, NY, USA, 2013, pp. 87–98.

[29] A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, F. Roli, Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning, ACM Computing Surveys (2023). URL: https://doi.org/10.1145/3585385. doi:10.1145/3585385, just Accepted.

[30] A. E. Cinà, S. Vascon, A. Demontis, B. Biggio, F. Roli, M. Pelillo, The Hammer and the Nut: Is Bilevel Optimization Really Needed to Poison Linear Classifiers?, in: IJCNN 2021, Shenzhen, China, July 18-22, 2021, IEEE, 2021, pp. 1–8. URL: https://doi.org/10.1109/IJCNN52387.2021.9533557. doi:10.1109/IJCNN52387.2021.9533557.

[31] M. Pintor, L. Demetrio, A. Sotgiu, M. Melis, A. Demontis, B. Biggio, secml: Secure and explainable machine learning in Python, SoftwareX 18 (2022) 101095. URL: https://www.sciencedirect.com/science/article/pii/S2352711022000656. doi:10.1016/j.softx.2022.101095.

[32] L. Demetrio, B. Biggio, Secml-malware: Pentesting windows malware classifiers with adversarial exemples in python, arXiv preprint arXiv:2104.12848 (2021).