Putting Adversarial Machine Learning to the Test: Towards AI Threat Modelling

Ronan Hamon [*1] **Henrik Junklewitz** [*1]

Abstract

Adversarial Machine Learning (AML) has emerged as a field of research studying the security of ML models. For this Blue Sky idea, we argue that with starting regulatory activity on AI and the need to secure operational ML-based systems against external threats, the AML community should increase its part in this endeavour, and scale up the efforts of discussing and prototyping more realistic cases to establish a clear connection to actual cybersecurity practices.

1. Secure-compliant AI systems

Artificial intelligence (AI) has become one of the focal points of the ongoing digital transformation. As with any technology, proper regulation and standardisation are eventually needed to ensure that its use will stay safe, secure and respectful of fundamental rights, societal values and law (High Level Expert Group on Artificial Intelligence, 2019; McFadden et al., 2021). Thus, regulatory bodies and governments are already advancing respective digital policy agendas, with the proposal for a regulation of AI (so-called AI Act) (European Commission, 2021) published in 2021 by the European Commission one of the major policy developments. In this context, a wide range of activities in different fields ranging from academic AI research to industrial software engineering have started to test various aspects of machine learning based software (Zhang et al., 2020), such as the robustness, interpretability and fairness. Among these aspects, this paper focuses on cybersecurity requirements, and discusses the relevance of adversarial machine learning (AML) for securing real-world software systems and ensuring conformity with regulatory requirements. It aims to highlight technical gaps between high-level legal requirements, technical methodologies, cybersecurity practices and techniques from AML (Berghoff et al., 2021).

[*]Equal contribution [1]European Commission, Joint Research Centre, Ispra, Italy. Correspondence to: Henrik Junklewitz <henrik.junklewitz@ec.europa.eu>.

AdvML Frontiers workshop at 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, 2022. Copyright 2022 by the author(s).

AML topics are getting very relevant in the task of practically securing AI systems. On the one hand, it is clear that even though AML has often been motivated by security problems, most of the literature has been focused on studying a set of problems important for fundamental questions of ML robustness and generalizability (Szegedy et al., 2014; Gilmer et al., 2019). So far, these studies not always translate well to help solving real-world cybersecurity problems (Biggio et al., 2018; Gilmer et al., 2018; Carlini et al., 2019). For example, many works on adversarial examples rely on restricted threat models based on constrained optimization (e.g., L_p-norm based adversarial attacks aiming enforcing low-intensity perturbations). This provides valuable insights about the functioning, accuracy and reliability of models, but it has been argued that the specific type of threat models considered are of limited use when connected to real-world problems.

On the other hand, connecting the current AML tools and threat models more tightly to realistic cybersecurity procedures are necessary to leverage the accumulated academic knowledge for daily cybersecurity practice. Usually, concrete threats for deployed software systems are analysed following a very applied and system-specific risk analysis framework (Shostack, 2014), which is much broader in scope than in AML. To this date, studying this more applied approach of modelling threats in AML remains an underrepresented field of study (Gilmer et al., 2018; Croce et al., 2021), especially for complex deep models and/or in cyber-physical contexts. It has already been pointed out that practical security of ML software systems needs to be developed with a dedicated approach (Huang et al., 2011; Papernot et al., 2018). As a matter of fact, many technical challenges considered as core components to any cybersecurity conformity testing with regulatory requirements remain open questions, such as the feasibility of measuring robustness against cyberattacks on ML models (Zhang et al., 2020; Carlini et al., 2019), or properly assessing the strength of defences (Tramer et al., 2020).

2. Towards AI Threat modelling

A number of potential directions have been identified in the recent literate to successfully harness practices and knowledge of AML and develop organisational and technical tools

for AI threat modelling and applied cybersecurity. Work could be invested into *broadening threat AML models* (Papernot et al., 2018; Gilmer et al., 2018; Biggio et al., 2018; Carlini et al., 2019), including for instance assumptions on realistic attacker's goals and resources, and metrics to measure attack success and robustness of systems in the environment in which the AI system evolves. The development of concrete *threat scenarios based on realistic applications* may help assessing the relevance of AML methodologies in operational contexts (Sommer et al., 2010). This is particularly relevant for AI systems as their low level of deployment largely requires to anticipate the threats and technical gaps in relatively new AI technologies. For such scenarios, it also needs to be taken into account that attackers in real world systems are likely seeking to minimize their costs, which implies that AML threat models are only practically important if the attacker cannot employ more simple means to achieve similar results (Gilmer et al., 2018). *Current cybersecurity practices* could be fostered to become a point of interest for AML practitioners. Cybersecurity communities on their side have already moved into the security of AI in digital systems, and many initiatives have started to adapt existing tools to AI such as the systematic listing of AI vulnerabilities (The MITRE Corporation, 2022). As a crucial practice, since AI systems are typically embedded in larger digital systems, considering the threats to the system as a whole is essential for AML (Papernot, 2018).

We argue that advances in cybersecurity testing of AI systems will have to address these gaps, starting by discussing how to foster lines of research that directly incorporate and appeal to both sides. Often the existing literature is too academic for cybersecurity experts, while at the same time many of the sources on cybersecurity practices can be hard to relate to for ML researchers without cybersecurity background. The directions outlined are possible paths to narrow down the gap between AML and applied cybersecurity of ML systems. AML as a field of research cannot, on its own, provide all solutions needed to secure real ML-based systems, but we argue that the AML community should take these proposals up and get more active in this endeavor, scale up the efforts of discussing, analyzing and prototyping more real and realistic cases to establish a clear connection to actual cybersecurity practices. International regulatory actors and standardization bodies will inevitably advance to set up regimes of certification and auditing for AI, and the research community should play an active role to influence the outcome of these activities for the better.

References

Berghoff, C. et al. Towards Auditable AI Systems. Technical report, 2021.

Biggio, B. et al. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84: 317–331, 2018.

Carlini, N. et al. On Evaluating Adversarial Robustness. Technical Report arXiv: 1902.06705, 2019.

Croce, F. et al. RobustBench: A standardized adversarial robustness benchmark. Preprint arXiv:2010.09670, 2021.

European Commission. Proposal for a Regulation laying down harmonised rules on Artificial Intelligence, 2021.

Gilmer, J. et al. Motivating the Rules of the Game for Adversarial Example Research. Preprint arXiv: 1807.06732, 2018.

Gilmer, J. et al. Adversarial Examples Are a Natural Consequence of Test Error in Noise. In *Proceedings of the International Conference on Machine Learning*, pp. 2280–2289, 2019.

High Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI, 2019.

Huang, L. et al. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, pp. 43–58, 2011.

McFadden, M. et al. Harmonising Artificial Intelligence:. Technical report, Oxford Information Labs, 2021.

Papernot, N. A Marauder's Map of Security and Privacy in Machine Learning. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security*, 2018.

Papernot, N. et al. SoK: Security and privacy in machine learning. In *Proceedings of the IEEE European Symposium on Security and Privacy*, pp. 399–414, 2018.

Shostack, A. *Threat Modeling: Designing for Security*. 1st edition, 2014.

Sommer, R. et al. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *Proceedings of the 2010 IEEE Symposium on Security and Privacy*, pp. 305–316, 2010.

Szegedy, C. et al. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 2014.

The MITRE Corporation. MITRE ATLAS. https://atlas.mitre.org/, 2022.

Tramer, F. et al. On Adaptive Attacks to Adversarial Example Defenses. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1633–1645, 2020.

Zhang, J. M. et al. Machine Learning Testing: Survey, Landscapes and Horizons. *IEEE Transactions on Software Engineering*, pp. 1–1, 2020.