Adversarial Machine Learning Awareness: Designing a Video-based Intervention for Practitioners

Hanna Algedri, Benjamin Berens, Maximilian Noppel, Christian Wressnegger, and Melanie Volkamer
KASTEL Security Research Labs, Karlsruhe Institute of Technology

Abstract

Practitioners often lack the awareness of attacks against ML-based systems and the knowledge to protect against them. We present (as work-in-progress) a video-based awareness intervention to correct common misconceptions of ML practitioners. Our intervention is designed in an multidisciplinary collaboration and reviewed for clarity by ML practitioners. The video follows a storytelling approach in which an adversary attempts to break machine learning systems. We use a production line as a visual metaphor for the ML life cycle to showcase the different attack types. Before progressing to evaluate the intervention with ML practitioners, we ask the usable-security community for feedback.

1 Introduction

The security of machine learning (ML) is of critical importance as ML has become a foundational technology in various application domains, such as healthcare [11], mental health [4], and autonomous vehicles [9]. Studies reveal widespread uncertainty among practitioners about the attack surface of ML systems, possible mitigations against attacks, and the responsibility for securing intelligent systems [1, 2, 6, 7]. This circumstance may be based on the fact that ML systems pose different security challenges than traditional computer applications [3]. These differences led to a new research discipline, *Adversarial Machine Learning* (AML), but practitioners often lack the knowledge about these attacks and the training to correctly assess the risks coming with them [1, 2, 6, 7]. To bridge the knowledge gap and raise awareness among ML practitioners regarding adversarial threats, we develop a video-based intervention targeting attacks against machine learning systems. We deliberately restrict our scope to security attacks rather than privacy to avoid packing too much information into a short video. The development is grounded in the Mossano framework for creating security awareness measures [8], which prescribes a system sequence of design decisions. We outline these design decisions related to key stages of the framework.

USENIX Symposium on Usable Privacy and Security (SOUPS) 2026.
August 23–26, 2026, Hannover, Germany.

2 Preparation

Our awareness intervention targets machine-learning practitioners, aiming to raise their interest and encourage them to learn more about attacks on and mitigations for machine learning systems. ML practitioners are a comparatively understudied group in security research, many of whom have limited knowledge of model hardening and attack mitigation [1, 2, 6, 7]. The format of our intervention accommodates their limited time availability and is presented as a short-video, under 10 minutes in length. Videos can communicate complex information while engaging multiple senses [10] combined with online availability and scalability.

3 Attack Techniques

Practitioners frequently dismiss educational resources as impractical due to tight schedules and some mention that learning AML fundamentals would consume more time than they can spare [7]. The selection of security attacks covered in the video is driven by the interplay between practitioners' time constraints and the severity of the attacks in broader operational context.

Related work [1, 2, 6, 7] shows that AML awareness among practitioners is low. Therefore, our intervention starts with the basics and explains how attackers can subvert ML systems. The ML life cycle is traditionally divided into two phases, both being subject to attacks: (1) Training-time attacks target

the model during the learning phase [12]. For instance, in *data-poisoning* attacks the adversary either injects malicious samples into the training dataset or modifies existing samples. (2) Inference-time attacks, in turn, target a model that is already trained and deployed. As the latter attacks are more practical, as no access to training data and process is necessary, they form the primary focus of our intervention. Evasion attacks or adversarial examples are a class in inference-time attacks in which attackers craft adversarial inputs that cause the model to misclassify the attack sample (see Figure 1 for our presentation of an Evasion attack).

4 Mitigations

The intervention is also intended to provide the target group with examples how to prevent the types of attacks addressed. Possible mitigations for data poisoning attacks include verifying training data and model provenance, maintaining checksums and continuously monitoring the training pipeline. For evasion attacks, mitigations center on model robustness: training models on adversarial samples so they learn to recognize adversarial patterns, retraining promptly as new evasion techniques emerge, using ensemble learning to combine several models into a joint prediction rather than relying on a single model, as well as filtering and flagging suspicious incoming interactions for human review or smoothing prediction outputs to prevent sensitive information from being exposed [2].

5 Misconceptions

Beyond gaps in technical knowledge revealed by related work, ML practitioners have misconceptions about who is responsible to mitigate AML threats and frequently deflect the responsibility to dedicated security teams [7]. Practitioners also underestimate the feasibility of attacks. For instance, they incorrectly assume that effective evasion requires backend access [7] and they hold inaccurate beliefs about the prevalence of AML threats and the reliability of existing defenses [5, 7]. The existence of real-world AML is sometimes doubted, with practitioners questioning whether adversaries would gain anything from exploitation [1, 5, 7]. Further, some practitioners exclude AML considerations by arguing that they do not handle sensitive data [2]. Others conflate AML-specific threats with general cybersecurity concerns or consider them irrelevant to their work [1].

6 Design of the Measure

We design our video intervention by integrating several key components including the attack techniques to cover, relevant mitigation strategies, common misconceptions about both attacks and defenses, keeping in mind the target audience's existing knowledge, awareness levels, and practical time constraints.

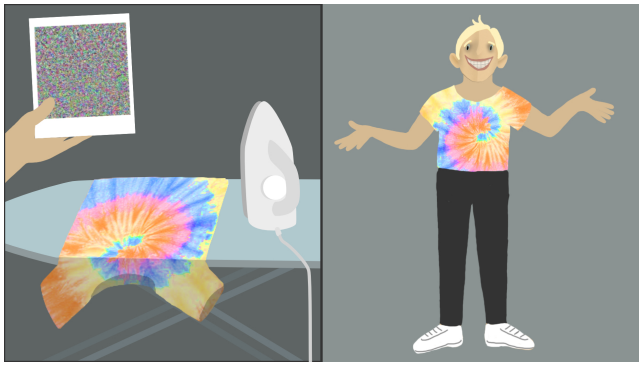

Figure 1: Example for an evasion attack. The adversary puts a pattern on a batik shirt which leads to a misclassification.

To develop the video, we work iteratively across an interdisciplinary research team spanning psychology, usable security and AML. We used a storytelling approach for designing the content of the intervention. Building on earlier attack-scenario videos from our research group [1], we craft a story [2] that begins by establishing artificial intelligence as a pervasive technology. An adversary identifies and demonstrates attacks on ML systems, succeeding against vulnerable systems and fails against systems that are already successfully implementing mitigation strategies. Visually, the video uses a production-line metaphor for the ML life cycle, serving as a recurring element that reinforces where attacks can occur. By mapping the life cycle alongside its vulnerable points, this metaphor also gives the audience a common language and knowledge base to build on. We keep the language accessible, favoring clarity over technical terms, so as to focus on essential concepts rather than expand the scope to additional attack types.

We recognize that a video cannot replace formal training. But it can build general awareness for AML-threats, spark curiosity, and motivate practitioners to pursue more learning or to discuss the topic with peers.

7 Conclusion and Future Work

While machine learning has become a core technology of our society, ML practitioners are insufficiently aware of its vulnerability in adversarial environments. In this poster, we present a security awareness intervention on the basis of a short video. This video has been verified by ML practitioners for its clarity, technical correctness, and narrative coherence. In the future we want to evaluate the effectiveness of our video intervention in the context of user study with ML practitioners.

[1] See https://www.youtube.com/watch?v=v6cq70RR_lc

[2] See our script here: https://osf.io/7f58g/overview?view_only=287eb038c41d46b991b9f2e242725ed8

Acknowledgment

This work was supported by funding from the project "Engineering Secure Systems" of the Helmholtz Association (HGF) [topic 46.23.01 Methods for Engineering Secure Systems] and by KASTEL Security Research Lab.

References

[1] Lukas Bieringer, Kathrin Grosse, Michael Backes, Battista Biggio, and Katharina Krombholz. Industrial practitioners' mental models of adversarial machine learning. In *Eighteenth Symposium on Usable Privacy and Security (SOUPS 2022)*, pages 97–116, 2022.

[2] Franziska Boenisch, Verena Battis, Nicolas Buchmann, and Maija Poikela. "i never thought about securing my machine learning systems": A study of security and privacy awareness of machine learning practitioners. *Mensch und Computer 2021*, pages 520–546, 2021. doi: 10.1145/3473856.3473869.

[3] Huaming Chen and M. Ali Babar. Security for machine learning-based software systems: A survey of threats, practices, and challenges. *ACM Comput. Surv.*, 56(6), February 2024. ISSN 0360-0300. doi: 10.1145/3638531. URL https://doi.org/10.1145/3638531.

[4] Simon D'Alfonso. Ai in mental health. *Current Opinion in Psychology*, 36:112–117, 2020. ISSN 2352-250X. doi: https://doi.org/10.1016/j.copsyc.2020.04.005. URL https://www.sciencedirect.com/science/article/pii/S2352250X2030049X. Cyberpsychology.

[5] Kathrin Grosse, Lukas Bieringer, Tarek R. Besold, Battista Biggio, and Katharina Krombholz. Machine Learning Security in Industry: A Quantitative Survey. *IEEE Transactions on Information Forensics and Security*, 18:1749–1762, 2023. ISSN 1556-6013, 1556-6021. doi: 10.1109/TIFS.2023.3251842. URL https://ieeexplore.ieee.org/document/10057473/.

[6] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *2020 IEEE security and privacy workshops (SPW)*, pages 69–75. IEEE, 2020.

[7] Jaron Mink, Harjot Kaur, Juliane Schmüser, Sascha Fahl, and Yasemin Acar. "security is not my field,i'm a stats guy": A qualitative root cause analysis of barriers to adversarial machine learning defenses in industry. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 3763–3780, 2023.

[8] Mattia Mossano, Anne Hennig, Fabian Lucas Ballreich, Benjamin Maximilian Berens, Filipo Sharevski, Martina Angela Sasse, and Melanie Volkamer. A framework for developing information security awareness measures. Poster presented at Twenty-First Symposium on Usable Privacy and Security (SOUPS 2025), Seattle, WA, USA, August 10–12, 2025, 2025.

[9] Alexandre Moreira Nascimento, Lucio Flavio Vismari, Caroline Bianca Santos Tancredi Molina, Paulo Sergio Cugnasca, Joao Batista Camargo, Jorge Rady de Almeida, Rafia Inam, Elena Fersman, Maria Valeria Marquezini, and Alberto Yukinobu Hata. A systematic literature review about the impact of artificial intelligence on autonomous vehicle safety. *IEEE Transactions on Intelligent Transportation Systems*, 21(12): 4928–4946, 2019.

[10] Michael Noetel, Shantell Griffith, Oscar Delaney, Taren Sanders, Philip Parker, Borja del Pozo Cruz, and Chris Lonsdale. Video improves learning in higher education: A systematic review. *Review of Educational Research*, 91(2):204–236, 2021. doi: 10.3102/0034654321990713. URL https://doi.org/10.3102/0034654321990713.

[11] Mohammed Yousef Shaheen. Applications of artificial intelligence (ai) in healthcare: A review. *ScienceOpen Preprints*, 2021.

[12] Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Andersen. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2024.