Mental Models of Adversarial Machine Learning

Lukas Bieringer[†,*], Kathrin Grosse[‡,*], Michael Backes[§], Katharina Krombholz[§]
[†]QuantPi,[‡]University of Cagliari,[§]CISPA Helmholtz Center for Information Security
[†]lukas.bieringer@quantpi.com [‡]kathrin.grosse@unica.it

Abstract

Although machine learning (ML) is widely used in practice, little is known about practitioners' actual understanding of potential security challenges. In this work, we close this substantial gap and contribute a qualitative study focusing on developers' mental models of the ML pipeline and potentially vulnerable components. Studying mental models has helped in other security fields to discover root causes or improve risk communication. Our study reveals four characteristic ranges in mental models of industrial practitioners. In this short abstract, we focus on the first range that covers the intertwined relationship of adversarial machine learning (AML) and classical security. Our work has implications for the integration of AML into workflows, security enhancing tools for practitioners, and creating appropriate regulatory frameworks for AML.

1. Introduction

Adversarial machine learning (AML) studies the security of learning based systems. For example, tampering with some features often suffices to change the classifier's outputs to a class chosen by the adversary [5, 2, 14] . Analogously, slightly altering the training data enables the attacker to decrease performance of the classifier [11]. Most attacks and mitigations studied in AML are in an ongoing arms race [1, 12, 4].

Although machine learning (ML) is increasingly used in industry, little is known about AML in practice. To tackle this question, we conduct a first study to explore mental models of AML. Mental models are relatively enduring, internal conceptual representations of external systems that originated in cognitive science [8]. In other security related areas, correct mental models have been found to ease the communication of security warnings [3] or enable users to implement security best-practices [15]. Mental models also serve to enable better interactions with a given system [16], or to design better user interfaces [6].

Our methodology builds upon these previous works by using qualitative methods to investigate the perception of vulnerabilities in ML applications. Our findings shed light on four characteristic ranges of practitioners' mental models of AML. In this short abstract, we describe the range that concerns the separation of AML and standard security. In many cases, the borders between these two fields are blurry: a subject may start talking about evasion and finish the sentence with a reference to cryptographic keys. On the other hand, security threats are often taken for granted, whereas practitioners are less aware of AML attack scenarios. The three other characteristic ranges cover functional and structural components, individual variations of perceptions, and differences in technical depth.

We found evidence that semi-automated fraud on ML systems takes place in the wild. Our findings on mental models allow to tackle these threats by (**I**) understanding how practitioners perceive AML threats, (**II**) developing tools that help to assess and evaluate security of ML applications, and (**III**) drafting adequate regulations that reduce insecurity about AML. Yet, more work is needed to understand the individual and shared mental models of practitioners.

Related work. Although AML research has been criticized for the limited practical relevance of its threat models [7], there is little work about AML in practice. To the best of our knowledge, only Kumar et al. [13] have studied this question. They found that practitioners are most concerned about poisoning and model theft and put their results in relation to the Secure Development Lifecycle developed for software.

2. Methodology

We describe the design of our semi-structured interview study, the drawing task, our recruiting strategy, the participants, and how we analyzed the data before we describe the results of our study.

*First two authors contributed equally.

Study design and procedure To assess participants' perceptions, we conducted semi-structured interviews enriched with drawing tasks, similar to Wu et al. [17]. The structure of our interviews covered the underlying ML pipeline of a given ML project a subject was involved in and possible security threats within the project. The detailed interview guideline can be found in a long version of this paper. As a last step in our interviews, we confronted the subjects with exemplary attacker models for some of the threats considered relevant in industrial application of ML [13]. To assess practitioners' understandings of these threats, the participants had to discuss the attacks within their setting.

At the beginning of the interview, participants were informed about the purpose of our study and the applied privacy measures. Each interview lasted approximately 40 minutes and has been conducted jointly by the first two authors of this abstract. Due to COVID-19, the interviews were remote and relied on a freely available digital whiteboard[1] for the drawing task.

Recruitment Recruitment for a study on applied ML in corporate environments is challenging, as only a small proportion of the overall population works with ML. We tried to initiate interviews with two multinational companies, however, both denied our request after internal risk assessments. We thus focused on smaller companies where we could reach out directly to decision-makers and convince them to participate. We relied on the individual networks of the authors and public databases[2], and used LinkedIn and emails to get in contact with potential subjects. Our requirements for subjects were a background in ML or computer science and positions such as data scientists, software engineers, product managers, or tech leads. We did not require any prior knowledge in security. After 15 recruited subjects, the research team agreed that the interviews saturated, and we stopped recruiting.

Participants Subjects were randomly assigned IDs for the study. One subject did not hand in the questionnaire and is thus not included in this description.

13 participants identified as male, one identified as female. As intended for a first exploration of practitioners' perception of AML, our sample covered various application domains and organizational roles: The companies' application areas were as diverse as healthcare, security, human resources, and others. Their subjects' roles were also diverse: Most subjects (8 of 15) were in managing positions. Three were software or ML engineers, three more were researchers. One subject did not report his role. Four subjects worked in companies with less than ten employees, five in companies with less than 50 and the remaining six subjects in companies with less than 200 employees.

Data analysis Our analysis adopted an inductive approach following recent work in usable security [10]. To distill patterns in interview transcripts and drawings, we applied two rounds of open coding. We then performed descriptive axial coding to group our data into categories and selective coding to relate these categories to our research questions. Throughout the coding process, we used analytic memos to keep track of emerging topics. The final set of codes can be found in a long version of this paper.

We calculated Cohen's kappa to measure the agreement among the coders. For drawings, we reached $\kappa = 0.85$, and for interview transcripts $\kappa = 0.71$. These values indicate a good level of coding agreement since both values are greater than 0.61 [9]. Given the semi-technical nature of our codebook, we consider these values as substantial inter-coder agreement.

Ethical considerations The ethical review board of our university approved our study design. We limited the collection of person-related data as much as possible. and complied with both local privacy regulations and the general data protection regulation (GDPR).

3. Perception of AML and security

In our interviews, the boundary between classical security and AML often appeared blurry or unclear, with the corresponding concepts intertwined. On the other hand, there were crucial differences in the perception between classical security and AML threats. One difference is that whereas security defenses were often clearly stated as such, AML mitigations[3] were applied without security incentives. Finally, we find a tendency to not believe in AML threats. Many subjects denied responsibility, doubted an attacker would benefit, or stated the attack does not exist in the wild. There was no such tendency in classical security.

3.1. Mingling AML and security

We first describe the intertwined perception of AML and security before we investigate if security and AML are used interchangeably using co-occurrence of codes.

Blurry boundaries. There are plenty of examples on vagueness about the boundary between classical security and AML. For example *S20* reasoned about evasion: "*this would require someone to exactly know how*

[1] https://awwapp.com/

[2] For example https://www.crunchbase.com/

[3] AML is far from being solved which we communicated to our subjects if required. Here, we define defenses as techniques which increase the difficulty for an attacker.

we deploy, right? and, where we deploy to, and which keys we use". At the beginning, the scenario seems unclear, but the reference to (cryptographic) keys shows that the subject has moved to classical security. Analogously, when *S18* reasoned about membership inference: "*but that could be only if you break in [...] if you login in to our computer and then do some data manipulation*". This subject was reasoning about physical access control as opposed to an AML attack via an API. Sometimes, ambiguity in naming confused our subjects. For example, *S11* thought aloud: "*poisoning [...] the only way to install a backdoor into our models would be that we use python modules that are somewhat wicked or have a backdoor*". Here, the term 'backdoor' in our interview triggered a standard security mindset involving libraries in contrary to our original intention to query subjects about neural network backdoors. Finally, *S12* stated: "*maybe the poisoning will be for the neural network. From our point of view you would have to get through the Google cloud infrastructure*". From an AML perspective, the infrastructure is irrelevant, as the model is independent. Yet, the infrastructure is perceived as an obstacle for the attack.

Co-occurrence of concepts. In the previous paragraph, we discussed the blurry boundaries between AML and classical security. Another example is *S6* reasoning about IP loss: "*we are very much concerned I'd say the models themselves and the training data we have that is a concern if people steal that would be bad*". In this case, it is left out how the attack is performed. Analogously, *S9* remarked: "*We could of course deploy our models on the Android phones but we don't want anybody to steal our models*". To investigate whether our subjects are more concerned about some property or feature (data, IP, the model functionality) than about how it is stolen or harmed, we examined the co-occurrence of AML and security codes that refer to similar properties in our interviews. For example, the codes 'model stealing' and 'code breach' both describe a potential loss of the model (albeit the security term is broader). Both codes occur together six times, with 'code breach' being tagged one additional time. Furthermore, the code 'model reverse engineering', listed only two times, occurs both times with both 'model stealing' and 'code breach'. However, not all cases are that clear. For example 'membership inference' and 'data breach' only occur together two times. The individual codes are more frequent, and were mentioned by three ('membership inference') and eleven ('data breach') participants. Analogously, attacks on availability (such as DDoS) in ML and classical security were only mentioned once together. Such attacks were brought up in an ML context twice, in classical security four times. Codes like 'evasion' and 'poisoning', in contrast, are not particularly related to any standard security concern. We conclude that AML and security are not interchangeable in our subjects' mental models to refer to attacks with a shared goal.

3.2. Differences between AML and security

We focus on the differences of the perception of AML and security by discussing defenses and threats and conclude with the practical relevance of AML.

Defenses. All fifteen interviewees mentioned a security defense (encryption, passwords, sand-boxing, etc). An AML mitigation appeared in eight. In contrast to security defenses, however, AML defenses were often implemented as part of the pipeline, and not seen in relation to security or AML. As an example, *S9*, *S15*, and *S18* reported to have humans in the loop, however not for defensive purposes. *S10* and *S16* were aware that this makes an attack more difficult. For example, *S16* stated: "*maybe this poisoning of the data [...] is potentially more possible. There, we would have to manually check the data itself. We don't [...] blindly trust feedback from the user*". Analogous observations hold techniques like explainable models (3 subjects apply, 1 on purpose) or retraining (2 apply, additional 2 as mitigation). For example, *S14* said: "*when we find high entropy in the confidences of the data [...] for those kind of specific ranges we send them back to the data sets to train a second version of the algorithm*". In this case, retraining was used to improve the algorithm, not as a mitigation. We conclude that albeit no definite solution to vulnerability exists, many techniques that increase the difficulty for an attacker are already implemented, however often unintentionally.

Perception of threats. There is also a huge difference in the perception of threats in AML and security. In security, threats were somewhat taken for granted. For example, *S9* was concerned about security of the server's passwords "*because anybody can reverse-engineer or sniff it or something*". Analogously, *S6* said to pay attention to "*the infrastructure so that means that the network the machines but also [...] libraries*". In contrast, almost a third of our subjects (4 of 15) externalized responsibility for AML threats. For example, *S1* said that ML security was a "*concern of the other teams*". Other reasons not to act include that subjects had not encountered an AML threat yet, concluding AML was not relevant. More concretely, *S9* remarked: "*we also have a community feature where people can upload images. And there could be some issues where people could try to upload not safe or try to get around something. But we have not observed that much yet. So it's not really*

a concern, poisoning". Roughly half of the subjects (7 of 15) reported to doubt the attackers' motivation or capabilities. For example, *S1* said: "*I have a hard time imagining right now in our use-cases what an attacker might gain from deploying such attacks*". *S20*, who worked in the medical domain, stated: "*I'm left thinking, like, why, what could you, achieve from that, by fooling our model*". Finally, many subjects (9 of 15) believed that they have defenses in place. As an in-depth evaluation of the effectiveness of defenses in each setting is beyond the scope of this abstract, we leave it for future work.

Practical relevance of AML. We might conclude that AML threats are academic—yet, our interviews showed that there are already attacks on ML in the wild. *S10* stated: "*What we found is [...] common criminals doing semi-automated fraud using gaps in the AI or the processes, but they probably don't know what AML [...] is and that they are doing that. So we have seen plenty of cases are intentional circumventions, we haven't quite seen like systematic scientific approaches to crime*". The unconcern of most of our subjects could then be an indicator that harmful AML attacks are (still) rare in practice.

4. Practical implications and limitations

We found that most our subjects lack an adequate and differentiated understanding to secure ML systems. Our findings can help to tackle these challenges by understanding which mental models exist and how to present adequate information to industrial practitioners. To this end, but also independently, our work helps to improve existing AML tools like libraries[4] and threat matrices[5]. Finally, we found that for many of our subjects, the European general data protection regulation served as a scaffold for their privacy perception. Similar frameworks for AML, which are on their way[6], could be supported by our work.

Our data is self-reported and subjected to a coding process. Yet, we continued coding and refining codes until a good level of inter-coder agreement was reached to limit the inherent subjective aspect of the data. With 15 participants, our sample size is rather small and limits generalizability. However, given the applied methods and that we reached saturation, the size is indeed acceptable [17]. Furthermore, AML is still evolving, thus the awareness of AML in the wild might increase. Our findings can thus only be valid temporarily. Finally, we studied a wide range of settings, each of which deserves a security analysis by itself, which is clearly beyond the scope of this abstract.

[4]For example the Adversarial Robustness Toolbox, CleverHans, RobustBench, or the SecML library, just to name a few.

[5]foe example https://github.com/mitre/advmlthreatmatrix

[6]https://ec.europa.eu/newsroom/dae/redirection/document/75788 for the EU draft on AI regulation.

5. Conclusion and future work

Based on our interviews, we take a first step towards a theory of mental models in AML. In this short paper, we focus on one of four characteristic ranges that concerns the relationship between AML and classical security. Both were often mingled, yet not used interchangeably by our subjects. For example, security and AML were not used interchangeably to refer to attacks with a shared goal. Security threats were also treated differently than AML threats: the latter were often considered less relevant. Finally our study provides evidence of variants of AML attacks in the wild.

A clear understanding of mental models of AML allows to improve information for practitioners, and helps to develop tools that assess the security of ML. Our work also highlights the need for regulatory frameworks that reduce uncertainty about and awareness of AML. However, a wide range of subsequent research towards an encompassing theory of mental models in AML is still required. This includes how mental models are shared, how classical security and AML are related, and work on threat specific taxonomies. Also AML tools and libraries could benefit from a clear understanding on how information should be presented and how these tools are used. Finally, we are convinced that the AML community will benefit from further practical assessment of attacks occurring in the wild, as our subjects only reported semi-automated fraud.

Acknowledgements

The authors would like to thank Battista Biggio, Antoine Gautier and Michael Schilling for their helpful feedback. This work was supported by the German Federal Ministry of Education and Research (BMBF) through funding for the Center for IT-Security, Privacy and Accountability (CISPA) (FKZ: 16KIS0753) and by BMK, BMDW and the Province of Upper Austria in the within the COMET programme managed by FFG in the COMET S3AI module.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks

against machine learning at test time. In *ECML PKDD*, pages 387–402, 2013.

[3] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, and Saranga Komanduri. Bridging the gap in computer security warnings: A mental model approach. *S&P*, 2010.

[4] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017.

[5] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *KDD*, 2004.

[6] Kevin Gallagher, Sameer Patil, and Nasir Memon. New me: Understanding expert and non-expert perceptions and usage of the tor anonymity network. In *SOUPS*, pages 385–398, 2017.

[7] Justin Gilmer, Ryan P Adams, Ian Goodfellow, David Andersen, and George E Dahl. Motivating the rules of the game for adversarial example research. *arXiv preprint arXiv:1807.06732*, 2018.

[8] P. N. Johnson-Laird. *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. 1986.

[9] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[10] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. Why do developers get password storage wrong? a qualitative usability study. In *CCS*, 2017.

[11] Benjamin IP Rubinstein, Blaine Nelson, Ling Huang, Anthony D Joseph, Shing-hon Lau, Satish Rao, Nina Taft, and J Doug Tygar. Antidote: understanding and defending against poisoning of anomaly detectors. In *SIGCOMM conference on Internet measurement*, pages 1–14, 2009.

[12] Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the effectiveness of low frequency perturbations. In *IJCAI*, pages 3389–3396, 2019.

[13] Ram Shankar Siva Kumar, Magnus Nystrom, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. *Available at SSRN 3532474*, 2020.

[14] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

[15] Madiha Tabassum, Tomasz Kosinski, and Heather Richter Lipford. " i don't own the data": End user perceptions of smart home device data practices and risks. In *SOUPS*, 2019.

[16] Rick Wash and Emilee Rader. Influencing mental models of security: a research agenda. In *Proceedings of the 2011 New Security Paradigms Workshop*, pages 57–66, 2011.

[17] Justin Wu and Daniel Zappala. When is a tree really a truck? exploring mental models of encryption. In *SOUPS*, 2018.