



# “Security is not my field, I’m a stats guy”: A Qualitative Root Cause Analysis of Barriers to Adversarial Machine Learning Defenses in Industry

Jaron Mink\*   
University of Illinois  
at Urbana-Champaign

Harjot Kaur\*  
Leibniz University Hannover

Juliane Schmäser\*   
CISPA Helmholtz Center  
for Information Security

Sascha Fahl   
CISPA Helmholtz Center  
for Information Security

Yasemin Acar  
Paderborn University &  
George Washington University

## Abstract

Adversarial machine learning (AML) has the potential to leak training data, force arbitrary classifications, and greatly degrade overall performance of machine learning models, all of which academics and companies alike consider as serious issues. Despite this, seminal work has found that most organizations insufficiently protect against such threats. While the lack of defenses to AML is most commonly attributed to missing knowledge, it is unknown why mitigations are unrealized in industry projects. To better understand the reasons behind the lack of deployed AML defenses, we conduct semi-structured interviews (n=21) with data scientists and data engineers to explore what barriers impede the effective implementation of such defenses. We find that practitioners’ ability to deploy defenses is hampered by three primary factors: a lack of institutional motivation and educational resources for these concepts, an inability to adequately assess their AML risk and make subsequent decisions, and organizational structures and goals that discourage implementation in favor of other objectives. We conclude by discussing practical recommendations for companies and practitioners to be made more aware of these risks, and better prepared to respond.

## 1 Introduction

Modern-day organizations use machine learning (ML) for various essential tasks from financial forecasting [27] to protecting their software systems [14]. With its growing popularity, ML has also become an appealing target for attackers. Research reveals that ML models are often vulnerable to adversarial attacks, resulting in data leaks, forced classifications, and degraded model performance. These threats have begun to materialize, with several organizations already becoming victims of AML attacks [12, 100]. In response, others have begun to advocate for securing ML systems via responsible AI guidelines in addition to traditional security measures [33, 53, 68]. Despite the abundance of research and importance of ML models in industry, prior work has shown that AML attacks

are widely misunderstood among ML professionals [13] and insufficiently defended against in organizations [78]. However, it is not well understood *why* these phenomena occur.

This paper provides a better understanding of why AML threats are not mitigated in industry despite active vulnerabilities and reported concern. While previous works have begun to explore industry perceptions regarding AML [13, 78], to the best of our knowledge, we are the first to report the barriers ML developers face when implementing AML defenses. To approach this problem, we conducted 21 semi-structured interviews with data scientists and data engineers (“ML practitioners”) to understand the barriers they face when learning about, assessing the risks of, and implementing defenses against AML threats in an organizational setting.

In this paper, we answer the following research questions:

- RQ1** [*Exposure*] What barriers prevent ML practitioners from adequately understanding AML attacks, and their corresponding risks and defenses?
- RQ2** [*Assessment*] What barriers prevent ML practitioners from adequately assessing the risk AML poses to their systems?
- RQ3** [*Implementation*] What barriers prevent ML practitioners from effectively implementing AML defenses in their systems?

From these questions, we make the following findings:

**Practitioners lack institutional motivation and effective material for understanding AML:** In order to keep up with their actively evolving field, practitioners constantly learn new information as directed by the project, compliance, and educational requirements demanded from them; however, AML concepts are nearly never included in these requirements. Instead, AML is learned from surreptitious encounters if at all. Furthermore, resources to learn about AML concepts are still largely constrained to research papers, making the learning of concepts difficult for practitioners. Perhaps due to this disconnect, we also find strong underlying beliefs from several practitioners that ML and security & privacy (S&P) are completely separate fields.

\*These authors contributed equally to this work.

**Practitioners lack the ability to assess their system for AML vulnerabilities and exploits:** While the evaluation and monitoring of software systems are well-understood by practitioners, these methods for adversarial testing do not extend to ML models. We find that nearly none of our participants performed or considered adversarial evaluation of their models before releasing them and none of our participants reported monitoring their models for adversarial threats. Additionally, several practitioners held misconceptions of AML, leading to misapplied threat models and overlooked vulnerabilities. Thus, practitioners were largely not able to determine if their systems were vulnerable to AML before deployment, or actively attacked during deployment.

**Company structure and goals discourage practitioners' effective implementation of AML defenses:** We find that company-induced pressures, organizational factors, and a lack of resources impede practitioners' ability to implement defenses. In addition to unclearly assigned responsibility for model defense, both the ML and security team, as well various ML teams working on different parts of the ML pipeline were isolated and lacked communication among each other. This broadly prevented the cooperation and initiative required for effective defense implementation. Additionally, participants lacked supporting resources to implement defenses and access to production-ready defenses, further increasing the difficulty of implementation.

Based on these findings we provide practical recommendations for companies and practitioners to be more aware of these risks and prepared to respond to threats if necessary.

## 2 Background

Below we provide a background of AML attacks, defenses, and real-world threats.

### 2.1 Adversarial Machine Learning

AML describes a set of techniques that attempt to leak information from, or degrade the performance of a ML model via malicious inputs [41, 42] or training data [20]. Attacks typically try to incur misclassifications on a provided input [5], and leak information about the models or training data [9, 87]. These attacks are not only effective but have proven robust enough to be transferable to models they were not explicitly trained on [73, 87]. Attacks can be grouped into three major categories: evasion, poisoning, and exploratory attacks. *Evasion attacks* attempt to avoid, or force, a classification by a ML system. By carefully perturbing a small number of influential features for an input, adversarial examples are generated that maximize the model's error while minimizing differences, resulting in misclassification [5]. *Poisoning attacks* manipulate training data in order to influence the behavior of the model. This can decrease accuracy or install a backdoor which allows to force a specific classification based on an attacker-chosen attribute [22]. *Exploratory attacks* corrupt the confidentiality

of ML models. Using black-box access, techniques can reconstruct the functionality of a model [87], determine if an input exists in a training dataset [75], and extract both low-resolution (general statistics) and high resolution (specific examples) information from the original training data set [9].

Since the discovery of adversarial attacks, a strong literature of techniques meant to robustify models against adversarially-crafted inputs has been produced. *Empirical defenses* were first proposed to protect against AML attacks. These defenses utilized heuristic-based methods that include transformations [36], detection of adversarial inputs [35, 48, 98], and adversarial retraining [31] among other techniques [65, 80]. These defenses protect against known techniques while maintaining similar, or improved levels of accuracy to an unmodified network. Unfortunately, it has since been discovered that empirical defenses do not hold up under an adaptive attacker that accounts for the implemented defense in their optimization function [10, 18, 19]. In an effort to overcome this problem, the research community's focus shifted to *certified defenses* which provide formal guarantees on a model's resilience, namely how much perturbation is required to change a classification to any other class [70, 95], how much information is leaked over a number of queries about the data [2], and how many poisoned data points are required to change a model's test loss [81]. While these defenses can provide formal guarantees, they are not without fault – certified defenses often result in decreases in accuracy, increases in complexity, and may require utilization of larger models [49, 88, 101, 102, 105]. Given these trade-offs, ML practitioners have multiple techniques available to harden their models; however, it is not yet known how industry developers perceive such threats, defensive trade-offs, and how this may ultimately impact the implementation of discovered defenses.

Other types of attacks which are not unique to ML applications, such as denial of service attacks, can also be applied to ML systems [76] and can be similarly mitigated via traditional security measures [5]. As these threats are not unique to ML, they are not within the scope of this paper.

### 2.2 Why AML is a Real Threat for Industry

While starting out as an academic inquiry, adversarial attacks on deployed machine learning system have not only been shown to exist in-the-wild but have been *actively exploited*. Real world attackers have since poisoned VirusTotal's malware detection system to misclassify benign files as malware [12], evaded Ant Group's facial recognition systems to impersonate victims in financial systems [100], and poisoned Microsoft's chatbot "Tay" to make discriminatory speech against twitter users [44]. Additionally, due to the difficulty of AML detection and relatively recent advocacy for ML auditing [62], these reports may under-estimate the true number of AML attacks. Indeed, this is corroborated by a number of

researchers who have reported instances when they could actively evade [6, 11, 17, 28, 54, 66, 74, 91, 99, 103], poison [47, 93], and exfiltrate [6, 23, 54, 66, 84, 91] production machine learning systems. Given the increasing importance of machine learning models and effectiveness of their associated attacks, we find it critical to understand what socio-technical barriers impede the adoption of effective defenses.

### 3 Related Work

This work is strongly informed by prior investigation into the security and privacy (S&P) perspectives of software developers and ML practitioners.

**Perspectives of Software Developers** Several prior works have studied software developers’ perceptions and behaviors around S&P-related concerns.

*Resources.* Acar et al. [4] investigated the impact that information sources had on code security and found that Android developers that were only allowed access to Stack Overflow wrote significantly less secure code as compared to those that either had access to the official Android documentation or books. Another work found that for cryptographic libraries, poor documentation and a lack of code examples negatively impacted code security [3]. Additionally, Li et al. [46] revealed that Android app developers rarely discussed privacy concerns in Android development subreddit unless triggered by external factors. Similarly in our work, we find that a lack of AML available resources may impact practitioners’ exposure to AML content and their ability to implement defenses.

*Organizational factors.* Assal et al. [8] examined the interaction between developers and software security processes via surveys. Their results highlight that a lack of organizational support for handling software security, may lead to insecure practices. Furthermore, a survey with app developers found that security expert involvement was a key factor leading to more frequent security updates [92]. Sadly, they also found that less than a quarter of the surveyed developers had access to security experts. In this work, we similarly identify a range of organizational barriers (several of which unique to ML practitioners) that impede the implementation of defenses in ML systems.

**Perspectives of Machine Learning Developers** Prior work has investigated how the habits and focuses of ML developers may result in negative security outcomes. Through interviews with AI practitioners, Sambasivan et al. [72] discovered that even in high-stakes domains, developers tend to focus less on data quality and cleaning, which were typically perceived as arduous and un-glamorous, and instead over-emphasized model building. The authors attribute this mentality in part to the under-emphasis of data engineering within degrees and required courses. As such, the authors recommend requiring data engineering courses along with data ethics and responsible AI for general AI degrees. Unfortunately, prior research into area-specific ethics of AI/ML courses [71] also noted

that practitioners may not naturally encounter such courses – while half of the analyzed universities offered an AI-specific ethics course, such courses were not always required by the degree and only a small number of technical courses had integrated ethics in the curriculum, possibly contributing to this lack of focus by ML practitioners. In our work, we extend this by finding that ML practitioners are also not typically exposed to S&P/AML concepts. Like ethics, this results in a lack of practitioner interest towards S&P/AML concerns.

**Perspectives on Adversarial Machine Learning** Despite the increasing research efforts and publications on AML attacks and defenses, relatively few works exist on understanding the ML practitioners’ perceptions and responses towards such threats. Kumar et al. [78] explore ML developers’ and security personnel’s perspectives by interviewing security and ML team leads in various companies. Overall, they found that most organizations lacked tools to protect their ML systems and required guidance to solve such issues. Bieringer et al. [13], interviewed industry practitioners and found that while ML developers might intermingle traditional and AML security, these topics were not used interchangeably. Furthermore, security threats are considered to be more relevant whereas AML specific mitigations appear only in some interviews and AML threat responsibility was externalized by some participants. In another work, Grosse et al. [34] conduct a quantitative study with ML industry practitioners to investigate the state of AML in practice and organizational AML aspects. In particular, they found that participants’ prior ML security knowledge had an influence on their attack concern and threat perception. In contrast to these works, we investigate a wide range of fundamental reasons to understand why practitioners lack exposure to AML knowledge, have difficulties in properly assessing AML risks, and run into barriers when trying to implement practical AML defenses. We then provide a set of recommendations that can help alleviate the cultural and organizational barriers that stand in the way of ML defense deployment.

## 4 Methodology

To investigate the various factors that contribute to a lack of deployed defenses, we interviewed 21 ML practitioners to discover their learning motivations, the risk assessments performed on their models, and organizational factors that influence the implementation of ML defenses.

### 4.1 Procedure

After an initial screening round to determine eligibility (Appendix A) we used an online survey<sup>1</sup> to collect participants’ professional and demographic data. We then invited eligible participants to schedule an online call. Interviews were conducted using a detailed interview guide to ensure consistency

<sup>1</sup>Due to space constraints, our interview guide, interview questions, and screening survey can be found in our replication package [55].

ID	Occupation	Experience	Application Domain	Data Used
P01	Data Scientist	1 year	Real Estate Value Prediction	Client Data
P02	Data Scientist	2 years	Document Summarizing	Client Contracts
P03	Data Scientist	3 years	Sales Forecasting & Product Recommendations	Internal Financial & Customer Data
P04	Data Scientist	3 years	Maintenance Image Classification	Customer Images
P05	Data Scientist	4 years	Financial Forecasting & Faulty Parts Detection for Automotive	Internal Financial/Production Data
P06	Data Scientist	4 years	Online Shopping & Advertisement Analysis	User Clickstream Data
P07	Data Scientist	6 years	Data Science for Defense Intelligence Agency	Classified Military Data
P08	Data Sci. & Manager	7 years	Value Predictions & Product Recommendations in Banking	Customer Personal & Financial Data
P09	Data Scientist	8 years	Distribution & Financial Forecasting for Postal Service	Customer & Internal Financial Data
P10	Research Scientist	13 years	Speech Recognition & Classification	Proprietary Audio Recordings
P11	Data Analyst	5 years	Macroeconomics Prediction for Government Finance	Country's Fiscal Data
P12	Data/ML Engineer	3 years	Telecom Data Streaming & Predicting	Client Telecommunication Data
P13	Data/ML Engineer	2 years	Scraping for Internet Media & Radar for Defense Contractor	Proprietary Government Data
P14	Data/ML Engineer	2 years	Classifying Oil and Gas Reserve Type	Synthetic & Proprietary Data
P15	Data/ML Engineer	2 years	<i>Did not disclose</i>	Client & Internal User Data
P16	Data/ML Engineer	4 years	Automated Transcription Service	Client Conference Call Audio
P17	Data/ML Engineer	14 years	Logistics Prediction/Text Analysis	Client Data
P18	Web Scraping Eng.	3 years	E-Commerce Monitoring & Predicting	Scraped E-Commerce Data
P19	SWE Intern	1 year	Geospatial Classification	Proprietary Government Data & Aerial Images
P20	Software Engineer	2 years	Classification of Surveyee Profiles	Internal Survey Participant Data
P21	Research Engineer	3 years	Sales Forecasting	Client Financial Data

Table 1: **Participant Experience & Domain.** A detailed overview of interviewees including their occupation, data science-related experience, application, and data they used. Application domain and data were determined via collaborative coding with three authors.

between interviewers [55]. The interview guide included advice for conducting qualitative interviews adapted from Rader et al. [69], the interview questions, and instructions regarding the procedure to be followed at the beginning and end of each interview. Whenever possible, a co-interviewer was present.

## 4.2 Interview Content

In our interviews, we sought to understand participants' perceptions of AML risks and how they interact with their learning habits, the structure of their organization, and their existing perceptions of S&P risks. To do so, our interviews consisted of four primary sections that we discussed with each participant. First, participants were asked to explain their data science background and learning habits (*Background and Education*). Second, participants provided us with an overview of their team's workflows and organizational structures, as well as their projects' goals and constraints (*Current Position and Project*). Third, participants discussed their perspective on general security issues, assigned responsibilities, procedures, and mitigations (*General Security*). Fourth, participants were asked about their knowledge and concerns regarding AML attacks and defenses, as well as any roadblocks they perceive towards effective defense implementation (*AML*). After asking if participants knew of AML and to explain their understanding of it, we provided an AML definition adapted from Bieringer et al. [13]) to ensure a common understanding for the remaining questions. In addition to these sections, the semi-structured nature of the interviews allowed us to explore participants' understanding, experiences, and opinions via follow-up questions [25]. We avoided priming participants for security or privacy when discussing their background and

current projects to see if either was brought up unprompted. We provide all questions in the replication package [55].

## 4.3 Data Analysis

We transcribed all interviews using a GDPR-compliant service. Three authors coded the transcripts in two coding rounds. The first coding round used a descriptive coding approach where each author labeled the broad interview concepts using topics derived from our research questions and interview guide. Each transcript was independently coded by two authors and subsequently verified via a meeting in which each disagreement was resolved. Disagreements were not due to genuine differences in interpretation but missed remarks, thus the codebook was not altered through this process. The second coding used an inductive approach where each author coded participant's thoughts in order to develop a set of themes. We iteratively discussed and refined codes in meetings until they stabilized. Similar to other work [24], we intentionally did not utilize inter-rater reliability (IRR) for several reasons [51]. First, the research goal was not to produce a set of codes or measure prevalence. Instead, we discover the existence of roadblocks that prevent AML-defense implementation. Second, IRR may lead to a simplification of information in an attempt to conform to a codebook. While our results are centered around emergent themes discovered through this process, the complete codebook can be found in our replication package [55].

## 4.4 Participant Recruitment & Demographics

To be eligible for our study, we required participants to be older than 18, English-speaking, and have at least one year



of experience developing or gathering data for predictive machine learning models in a company setting.

We recruited participants using the freelancing platform Upwork [1], where we offered the online survey and interview as a job and compensated participants \$89 for a 90-minute interview. We chose this approach as Upwork offers access to many freelancers specializing in data science [29] and has been successfully used by previous studies [40, 79, 90]. Furthermore, as Upwork allows us to view participants’ profiles and their uploaded CVs, we were able to ensure diversity of interviewed participants.

To determine eligibility, applicants filled in two screening questions describing their professional experience with ML, which an author then reviewed (see Appendix A). We piloted our interview with an acquainted ML practitioner, but since only minor changes were introduced, we include this pilot in our final data set. We invited 23 eligible participants to our study. Our goal was not to obtain a representative sample, but a diverse one with regard to their professional and cultural backgrounds. Despite the screening survey, 2 participants were found to not have the required experience during the interview and were subsequently removed. From 21 valid interviews, we had a total of 27 hours and 53 minutes of audio recordings with an average interview length of 80 minutes. We stopped interviewing upon reaching theoretical saturation, i.e., when no new themes emerged [25].

Our participants came from a variety of backgrounds. Demographically (Table 2), 17 participants identified as male and 4 as female, with ages ranging from 23 to 37. Participants were diverse, both geographically (located across 5 different continents) and ethnically (with white/European, South Asian, Middle Eastern, East Asian, Latino/a/x, and Southeast Asian representation). We also receive a variety of employment statuses and educational degrees. Likely due to methodological differences in recruiting, we received a lower number of participants with PhDs and many more with bachelor’s degrees than prior work [13]. Professionally (Table 1), participants reported between 1-14 years of working experience along with a variety of occupations, domain applications, and utilized data.

## 4.5 Ethics, Data Protection & Replicability

In our job offer, participants were informed of the study procedure and their compensation. Participants were given a consent form at the beginning of our online survey that informed them about data collection, data usage, and their rights. Before each interview, we reiterated participants’ rights and answered all questions to their satisfaction. Our entire process, including data collection, transcription provider, and data processing and storage, was GDPR compliant. Transcripts were anonymized by removing any identifying references before data analysis. This procedure was approved by our institution’s ethics review board as well as our data protection office.

In order to make our work reproducible, we provide all necessary supplementary material in a replication package [55].

## 4.6 Limitations

A number of limitations inherent to a qualitative interview approach apply to our study. This includes social desirability, confirmation, self-report, and recall biases. We mitigated these by carefully probing for answers and assuring participants that their answers would not be judged. Additionally, we asked follow-up questions to participants’ claimed AML knowledge to ensure correct interpretation<sup>2</sup> and depth of understanding. In order to minimize self-selection bias, we did not mention security or privacy in our recruitment process. While we achieved a diverse and international sample from different industries, due to the skewed demographics of ML practitioners on freelancing platforms [29], our participants skew towards young and male participants. Additionally, ML practitioners on Upwork may not be fully representative of all practitioners and generalizability cannot be assumed due to the limited sample size inherent in qualitative studies. We account for this by contextualizing our results appropriately (i.e., not generalizing).

## 5 Results

In our interviews, we find that nearly half of our participants did not know about AML and all but one had not implemented a defense. This is well-aligned with prior work which finds very little AML defense implementation among studied ML practitioners [13, 78]. Our results extend these findings by highlighting *why* this is the case, i.e., what barriers stand in the way of effective defense implementation against AML threats. First, to understand why there is a lack of AML understanding, we present reported challenges to AML exposure and learning. Second, to understand practitioners’ AML concerns and motivations (or lack thereof), we present reported challenges to effective AML risk assessments. Third, to understand why defenses are not put in place even when motivated, we present reported challenges to AML defense implementation.

### 5.1 Challenges to AML Exposure and Learning

To understand and mitigate AML vulnerabilities, ML practitioners must be meaningfully exposed to and taught about potential risks and mitigations. However, we find that a large number of our participants had not been exposed to AML concepts at all or had been taught in ways unfit for practical application. In particular, we find that participants were poorly exposed to useful AML resources due to a lack of

<sup>2</sup>Some participants conflated AML with Generative Adversarial Networks [26].

perceived relevance of S&P to ML concepts, AML requirements throughout their career, practical resources to learn AML from, and engagement with S&P-conscious teams.

**Practitioners Assume S&P is Irrelevant to ML.** Despite the ability of AML to affect the integrity of model predictions and confidentiality of data, several participants assumed that S&P concerns were irrelevant to their work in ML.

While many participants understood that training with proprietary or sensitive data could lead to privacy risks, these concerns never translated to their trained models: *“The main issue is about the data, the data should not get leaked. I don’t think anything specific to models, since in general, the data should not be leaked”* (P05). This assumption of irrelevance also came up implicitly when participants noted that S&P concerns were separate from their focus on ML. For instance, when asking P11 if they had ever encountered S&P topics in an ML course, they emphasized: *“No. I only focus on data analytics and predictions... it’s on how you build a neural networks... that’s my main focus.”* Indeed, after AML was defined in the interview, several participants noted that this is the first time they had even heard of this intersection: *“No, honestly speaking, I haven’t heard about the security issues at model levels, I’m aware of these things at data level, but at model level, this is the first time I’m hearing about these things”* (P02).

We also find that the assumption of irrelevance may have implications on how practitioners learn and mitigate AML risks. After hearing that unsecured models are vulnerable to AML threats, several participants expressed interest in learning about these concepts: *“It’s so important and it’s good to know about the security and the problems that models may face. But maybe I have to just search about that and find some courses to know more about different aspects of how people misuse the model”* (P14). Thus, simply explaining that there are S&P implications for practitioners’ decisions may encourage further learning of AML threats and their mitigations. However, other participants still felt this area was outside of their expertise and did not wish to pursue future knowledge: *“So honestly, I consider computer science and cybersecurity completely separate fields”* (P13). Additionally, this distinction may be used for justifying who is responsible (Section 5.3), which directly affects mitigations.

#### Key Insights:

- Several practitioners assume that ML and S&P are disconnected fields.
- This limits their learning and mitigation of AML risks.

**AML Knowledge is not Mandated by Projects or Institutions.** When asking participants how they learned new information, we found that requirements for project functionality, compliance trainings, and educational degrees were significant motivators for learning new knowledge. These re-

quirements were essential as several participants mentioned that the ML field has a constant output of new knowledge and only a subset can ever be learned: *“A lot is happening in the field; the new models for new data sets... it’s [a] huge [amount of] knowledge. Sometimes I think it’s a coincidence what you come across and what you [don’t]”* (P10). However, participants indicated that AML content was rarely covered in these requirements, leading to a lack of exposure.

**Task-based Requirements** and functional goals during model development were reported as a primary motivation for learning new content, but AML defenses were rarely part of these. Only one participant recalled AML being explicitly considered in an assigned task. As a defense contractor, P07 noted that their domain had a high amount of adversarial pressure, necessitating them to consider AML requirements more so than commercial sectors: *“[AML] concerns us, because we see it all the time. I understand you guys don’t always, because you might be concerned with different things... Facebook, for instance, is concerned a little bit more with its user base getting and staying on their platform than they are with results that go past that... [AML] sounds like a boogeyman to a lot of engineers.”* Furthermore, several participants noted that project requirements were often prescribed by external persons such as clients or management. However, these same participants also mentioned that such parties would likely not know about, or be concerned with AML, thus it may be unlikely for an AML requirement to arrive from them. Even in the defense domain, P07 noted that their clients often did not prescribe AML requirements. Instead, their team had to enforce them: *“A lot of the private data providers, their sales and their engineers have to be told like, ‘Hey, this is something that occurs.’ They’re like ‘oh yeah, I read a paper on that once. I heard that could occur.’ And it’s just like, ‘No, this is real.’”* Because this motivation to learn AML is not externally provided, teams who do not already know about AML may be less likely to learn.

**Compliance-based Requirements**, such as certification courses or enterprise trainings, was the sole source of exposure to S&P knowledge for multiple participants. However, no participant reported learning AML, or data-science focused S&P concepts in any such course. Instead, participants reported taking generic enterprise S&P trainings assigned by their company. These courses commonly contained information about safe password usage, safe access control on code/data, and phishing training: *“There’s usually like very, very basic stuff but aren’t specific to the tech domain where it’s like ‘hey, don’t write your passwords down on this sticky note and leave your desk’ or like ‘don’t text your friends like the last four digits of your Social Security number’, like very, very general things”* (P03).

A few participants took specialized security courses. These included commercially available courses such as CompTIA Security+, or domain/company-specific ones such as banking accreditation. Unlike the enterprise security courses, these

were often taken to comply with clients' or governmental agencies' required practices. When these certifications were required for business, they were prioritized by participants. P15 noted that they only took S&P courses because of clients' requirements: *"I did some of the certifications and stuff you do to convince the big clients that you are really data privacy oriented."* However, as noted in task-based requirements, participants felt as though external teams were unlikely to be aware or concerned of AML threats, thus it may be even more unlikely that clients would enforce AML-oriented certificates for hired ML practitioners. Furthermore, this issue may be compounded by the fact that there is a lack of commercial AML courses available. To the best of our knowledge, no such certification yet exists, meaning that clients are unable to require standardized AML compliance even if they wanted to. Instead, clients' currently have to provide their own training or be content with the teams' current AML posture.

**Educational Requirements** were a common way for participants to learn a wide breadth of knowledge required for their job. However, we find that the university and online programs taken by ML practitioners largely lack AML content in their required curriculum and courses. Only a few of our participants mentioned learning about AML through university coursework, all of which were in master's programs at the time. As 38% of ML practitioners hold only an undergraduate degree [39], this may imply that such knowledge is not often taught to a substantial portion of ML practitioners. Furthermore, given the many participants who took undergraduate and graduate courses yet did not see AML coverage, there exist many programs which do not require AML topics to be covered. To further complicate matters, even if AML concepts were required in university curriculums, these requirements may be unintentionally evaded for ML practitioners without a tech-related degree.

For many of our participants and non-tech students in particular, online programs, such as Coursera or Udemy, were also an important source of ML knowledge. However, none of our participants report encountering AML through an online learning program. Thus, a similar lack of AML coverage appears to exist in popular online ML courses.

#### Key Insights:

- Requirements from education and companies motivate practitioners' learning, but do not include AML topics.
- AML content may not be included as the stakeholders who create requirements may not be knowledgeable of AML threats, or AML content may not be viewed as required content towards tech-related degrees.

**AML Resources are Inapplicable or Unavailable.** Compared to resources for other ML concepts, AML resources are either available and difficult for practitioners to use, or likely unavailable.

**Academic Papers** were used by participants for learning ML and AML knowledge. However, participants also noted

that their learning was limited due to difficulties in exposure, comprehension, and translation to a real-world application.

A few participants mentioned that time constraints and the sheer amount of published research made it difficult to keep up date. P20 noted that having to read the papers needed to understand AML is a large time sink and a significant barrier: *"One of the roadblocks [for AML understanding is]... I don't have the time to read all research papers to be as up-to-date as a specialist in the area... So that's a real lack of knowledge, awareness would be a whole block."*

Even once a paper is found, several participants found academic papers difficult to comprehend. While P07 often read papers to gather new knowledge, they perceived that most practitioners aren't able to understand research papers: *"For every ten that I know, maybe one [practitioner] will be able to read the academic papers."* In addition to complex ideas, academics may assume a foundation of knowledge that practitioners may not have while solving a specific issue: *"[Academic papers] directly step inside the concept at high level... But for a person like me, who just know[s] there's something called 'X', I don't get other resources to study much about it. I feel they [should] take [it] from the basics: this 'X' is used, and where it is used, and why they are using it here. Some implication on the importance of the particular thing they're doing, instead of directly stepping inside... I feel I have to read it some three or four times, so that I can understand what's going on"* (P02).

Furthermore, several participants felt that it was difficult for them to effectively apply ideas from research papers in real world systems. Several participants noted that papers often struggled to present practical implications and use cases in their work: *"When I'm reading papers a lot of it will just be kind of, 'look at this cool rabbit hole that [academics] dug into and hyper-optimized for'... But rarely do I see kind of the impact beyond the scope of 'hey, look at this cool new technique'. Like not many people push it pass like the finish line of saying this is how it will impact the world, or the feasibility of using this in practice"* (P03). Several participants also reported that the lack of productionized software or packages to implement these tools further compound such issues, both increasing the difficulty to understand and implement novel concepts in production (see Section 5.3).

**Blogs and Entertainment** were used by participants to provide a broad, accessible overview of AML topics, but were perceived to lack actionable or insightful information. However, this accessibility also allows for greater reach of AML concepts compared to other resources.

While accessible, participants noted that blogs, podcasts, and news provide little in-depth knowledge on the mechanisms or technical details. As such, this media was often used as a first step towards a deeper understanding in a particular subject: *"A lot of [practitioners] start their journey looking at blog posts which are the highest-level dilution of ideas, and then from there, they'll dig deeper into whatever facet they're*

looking for” (P03). However, while this deeper progression was common in non-AML topics, no participant reported using blogs to start a deeper understanding of AML topics. While we cannot say what this is due to, we hypothesize it may be due to a lack of perceived relevance to their job, or a lack of actionable resources embedded in the media for further learning. Furthermore, these outlets were perceived as incompletely covering the topic, favoring attacks while leaving out remediations: *“I [learned about AML] indirectly, just in various podcasts. Basically, it was more high level. It didn’t go into as much detail on the modeling-side, more on the hacker-side or breaching-side”* (P08).

Because of their accessibility, entertainment media is able to reach more practitioners than more traditional resources such as papers and coursework. Several participants noted that their exposure to AML was solely provided through vectors like Medium, Twitter, Reddit, and various podcasts. These encounters were often coincidental rather than intentional: *“I’m just listening to like a cool podcast about people getting around image detection, or fake reviews for example. But not in the context of like my own work. It’s always in the back of my mind but without that sort of external random stuff I would have never heard about”* (P03). Thus, while incomplete, such entertainment media allows practitioners who wouldn’t normally care about AML to be exposed to these concepts. However, it is not obvious to what extent this results in meaningful protections.

**Educational Courses**, such as online and university courses, were reported by several participants to be essential in learning fundamental skills. However, for emerging topics like AML, courses seem either slow to incorporate new topics in traditional lectures, or act as paper reading groups rather than structured learning.

Several participants noted that structured courses did not keep up to date with active research in many fields. P02 noted this frustration: *“I take these courses and yes, those are the basics and I know them, but I don’t know what currently is going on in “X”... when we get these courses and I feel like that is something which is not updated.”* P20 corroborated this, noting that the slow production of some of these courses forced them to utilize other resources: *“In some areas - especially when they are so new - there are no tutorials, online videos or someone that’s going to learn you how the things work. In these cases, I’m first going to read some papers. There are especially, for example, some good blogs: ‘Towards Data Science’ or ‘Medium’”*. Two of the three participants that took AML university courses noted that it was heavily paper-based, perhaps running into the same issues previously mentioned. The other participant notes that they only discussed AML via *“two or three lectures regarding data privacy”* in a data science-based course. Unfortunately, this only led to light insights in such threats: *“I think [AML] was mentioned at universities at some of the lectures, but most of them did not go into details”*. However, even in these guided spaces, partic-

ipants did not come away with a holistic picture of threats and defenses. While all three participants could report at least one threat that AML posed to real systems, none of them were able to concretely describe a data- or model-defense to such threats. This may imply that like media, AML topics in courses are also skewed towards exploits rather than defenses. Perhaps because guided online courses require a concrete structure that is difficult to manage for quickly emerging fields, no participants mentioned taking an online course around AML (nor, to our knowledge, does any AML course currently exist on the popular platforms Udemy and Coursera).

**Other Resources** were found to be used for general data science and ML knowledge yet were not reported while learning AML. This may suggest a dearth of practical AML resources currently available to practitioners.

Among our participants, several made use of community forums, books, and existing code/data to learn basic data science and ML knowledge necessary for their job. However, no participants mentioned using any of these resources to learn AML topics. As participants did not comment on the reasons for not using particular resources, it is not known whether these resources weren’t readily available, not found through participants usual discovery habits, or absent for other reasons. Regardless, we see that several avenues of knowledge are not utilized for AML learning.

#### Key Insights:

- Hard to understand research papers and high-level entertainment media were the primary AML resources.
- Common resources for ML concepts were not used for learning AML, potentially implying unavailability.
- Emerging topics, like AML, were perceived as difficult to develop structured learning material for as they would quickly be outdated.

**Practitioners Don’t Interact with AML-knowledgeable Colleagues.** In their immediate network and company-sponsored events, practitioners are not exposed to AML.

Participants noted that close mentors and peers play a key role in affecting what topics they’re exposed to and what they learn. Colleagues recommended courses to take, papers to read, conferences to attend, new technologies to investigate, new features to add, and even how to learn. Furthermore, colleagues were often used as a resource to help others understand complex topics. For instance, several participants mentioned that colleagues and mentors were consulted by practitioners to help understand academic papers or how to implement a specific feature. As only P07 reported having access to colleagues that identified as “cybersecurity ML experts”, all other participants may not have a colleague to learn AML topics from. Several participants noted their colleagues likely had the same lack of exposure: *“[My colleagues have] the same view as mine now in the sense that we were not exposed to this area of research.”* (P19).



Furthermore, participants noted that this lack of AML knowledge was equally present in dedicated learning events held by their companies. From company-sponsored “hackathon” competitions to cross-team “lunch-and-learns”, participants reported that their companies often encouraged or required their data science division to participate in a number of social engagements. Generally, the goals of these events were to elevate the core competencies of teams and introduce them to knowledge outside of their immediate project. P03 perceived company-sponsored events, such as conferences, as especially influential for practitioners who don’t actively learn on their own: “*For the vast majority of data scientists, their new information comes from these conferences because the company pays for it.*” However, no participants reported AML topics or security team members participating in these events. For instance, P13 noted that their company wanted diverse ideas at their lunch-and-learns: “*If you had been working on something that, would not necessarily [be something] someone would care about or something like on an external team, you could then do a small presentation about it.*” However, when asked if a security-relevant talk ever occurred, they quickly reply: “*No.*”

#### Key Insights:

- Many practitioners do not interact with AML-knowledgeable colleagues in their immediate network or at company-sponsored events.
- Without this community, a common method for learning ideas is absent for AML.

## 5.2 Challenges to AML Risk Assessment

In order to decide on the protective actions needed to defend a system, ML practitioners first need to have a precise understanding of the vulnerabilities within the scoped systems. However, we find that practitioners often lack the ability to determine their risk against AML attacks. In particular, we find that practitioners struggle with risk assessment as they broadly do not evaluate models for AML risks, do not monitor their models for AML attacks, and hold misconceptions about AML threat models.

**Model Evaluations do not Account for AML.** Several participants routinely evaluated their models for performance, correctness, and accuracy. However, adversarial evaluation was scarce and AML vulnerabilities were nearly never accounted for.

While some participants interacted with security teams that inspected their ML systems and performed penetration testing on their infrastructure, only one participant reported proactively evaluating their model for AML risks. All other participants either focused on general security precautions such as endpoint/data pipelines security testing for non-AML issues or did not perform any security evaluations at all: “*In*

*the companies that I work, nobody cares about [ML security]” (P09).*

Although we find that automated model evaluation was ubiquitous among participants (and indeed a requirement to build and train the model in the first place), most of those methods were unsuitable for finding and recognizing adversarial vulnerabilities. When automatically evaluating their models, practitioners typically evaluated the performance via typical metrics including, but not limited to precision-recall, F1-score, and AUC-ROC curves. Though these metrics help evaluate and improve model performance characteristics such as accuracy, they fail to address model security and data privacy. These evaluations have no bearing on the model’s vulnerability to AML methods [16], thus leaving them unchecked.

Another subset of participants mentioned testing the model manually. While this may have the ability to identify if a model is vulnerable to AML attacks in specific scenarios, that did not appear to be the intention. Several participants mentioned manually testing their models for edge cases or unexpected inputs that cause misclassifications. For instance, P05 encouraged their team to try and break their model by providing any possible inputs: “*We do extensive testing. Basically, let’s say five of us are working together, I build some model, and I say, ‘Give the model any input you can’. And basically, I’ll say, ‘I’ll give you a treat if you can break the model’.*” In some cases, participants would consult with stakeholders outside of their team, and give the models to their clients to test and provide feedback. Thus, it’s possible that these participants got some sense of whether their models are vulnerable. This may imply that some practitioners do care about evaluating security properties of their models but lack the methods to perform them effectively. Similarly, others mentioned edge case tests that were focused on non-malicious errors such as testing for different formats, minus numbers, big numbers or complex sentences (in case of textual inputs), rather than adversarial conditions. Thus, they were unlikely to find vulnerabilities to more sophisticated attacks, gradient-based AML attacks.

We found mention of model-independent evaluations such as “code reviewing”. However, it wasn’t mentioned, and thus unlikely, that such reviews searched for AML vulnerabilities.

#### Key Insights:

- AML evaluations is not performed among several practitioners; thus, models may often be released without practitioners knowing how vulnerable they may be.
- Practitioners tend to use simple, manual heuristics that may indicate weaknesses but are not complete (e.g., edge-case testing).

**Model Monitoring is Unlikely to See AML Exploits.** Monitoring and visibility are important to ensure effective responses to security incidents. However, while some participants monitored their models for performance and main-

tenance, none of our participants reported monitoring their models for AML attacks.

Several participants used metrics (e.g., software engineering metrics or model accuracy) to monitor their models in production. Though not explicitly noted in the context of adversarial scenarios, typical monitoring metrics such as accuracy might help identify AML attacks indirectly. For instance, P15 mentioned the concept of "data threats": *"What we also try to do is we try to model this concept of data threats, whether the data is changing, like maybe we started with some other data now that the data is changing, and we are not noticing it. From time to time we do some statistical comparison between the data now versus one month ago."* Similarly, some participants relied on client or end-user feedback in case of errors as a model monitoring method: *"if the client is like saying that if he observe some errors, then we look into this and see what's going on"* (P10). Thus, AML attacks that significantly alter model performance (e.g., poisoning) or repeatedly interact with the model (e.g., exploratory attacks) may be detectable, attacks that require less interactions or are more subtle (e.g., evasion) may remain undiscovered. Furthermore, this detection would only after damage has already been occurred by harming both system functionality, and its users.

Lastly, several participants completely disregarded model monitoring, leaving them unaware of the status of production models.

#### Key Insights:

- Model monitoring is not ubiquitous among practitioners.
- Practitioners who do monitor their models, do not have the visibility to detect all possible AML attacks.
- Practitioners may not know whether models have been previously exploited or are undergoing active AML attacks.

**Possibilities for Overlooked Vulnerabilities.** Even when participants reported hearing and understanding AML concepts, in multiple cases this knowledge was found to be incomplete or incorrect which may lead to unresolved vulnerabilities.

Several participants held misconceptions of the required threat model for AML attacks, resulting in the possibility for overlooked vulnerabilities in their systems. For instance, P15 works in a data-science contracting company in which their models were made available via an exposed API. When asked if any AML attacks were concerning, they noted that adversarial examples were concerning, but only possible if the attacker had access to their backend, which was difficult and would result in other, larger concerns: *"I see maybe cases where [adversarial examples] can happen, but it can only happen on the backend of our company. Like if someone really managed to get so deep inside backend that they can also modify our data... For the current state, it shouldn't be very problematic."* However, evasion attacks require a much weaker threat

model than P15 assumed. Adversarial examples are able to be developed for black box systems without knowledge of the internal system [37,43,64], and thus, can likely be constructed for the exposed API endpoints *without* a need to access their backend. Thus, P15's system was unknowingly vulnerable to adversarial attacks due to a misconception in their perceived threat model of AML attacks.

In addition to incomplete threat models, we find that misconceptions related to effective AML defenses may lead to system vulnerabilities. In order to defend against AML, attack-specific mitigations may be necessary (e.g., randomized smoothing for evasion attacks [95], differentially private models for exploratory attacks [2]). However, some participants incorrectly believed that generic security measures were enough. P08 for instance worked in an organization in which sensitive internal financial data was used to predict property prices, credit scores, and other relevant data for clients. Because of this, clients were able to provide inputs and receive outputs trained ML models. While P08 was most concerned with data breaches, they noted that their current security protections should be enough for AML attacks: *"In terms of security, I think we are covered... If someone were to launch an attack, then [security team] would be a first line of defense before it comes to our space. We are not that concerned."* However, this sense of security was misguided. If a client is untrustworthy and has the ability to query a model, data can be exfiltrated unless protected by AML-specific techniques [2]. Thus, a misconceived notion of AML defenses may result in practitioners leaving their models vulnerable.

Furthermore, we find that it is easy for practitioners to develop misconceptions when first learning about AML. Towards the end of the interview, we provided participants with a definition and multiple examples of AML attacks based on prior work [13]. However, during this introduction, several of the aforementioned misconceptions materialized including incorrect understandings of both the attacker model and effective defenses. Only after a detailed discussion, were these misconceptions resolved. Thus, to ensure that practitioners do not develop misconceptions, it is pertinent to develop resources that clearly articulate when a threat is possible, and what defenses are (and are not) able to provide protection.

#### Key Insights:

- Misconceptions about AML and its defenses exist and might lead to unnoticed and unmitigated vulnerabilities.
- Learning resources must be carefully developed to avoid misunderstandings and focus on clear communication of applicable threats and mitigations.

### 5.3 Challenges to AML Defense Implementation

We find that there are several challenges for the implementation and application of defensive AML techniques in practice. Our interviews uncover organizational and implementation-

specific factors that impede practitioners' ability to defend against AML threats. These include the isolation of ML teams, undefined responsibility for AML defenses, competing business objectives, and a lack of knowledge about applicable defenses.

**Isolation of ML and S&P Teams Prevents AML Collaboration.** Isolation between different teams within the ML pipeline, as well as ML teams and S&P teams hinders the exchange of knowledge and increases the difficulty of implementing effective AML defenses.

ML systems uniquely blur the line between data and code, and AML attacks often span multiple parts of the ML pipeline (i.e., data collection and processing, model development and evaluation, and deployment) as the point of entry for an attack can be at a different stage than the effect of the attack. For example, data poisoning-based backdoor occurs in the data collection phase, must be defended against in the data collection or model development phase, while the resulting effect of a backdoored classification will only appear in production. Unfortunately, we find that teams working on different parts of the ML pipeline were isolated from each other, and lacked the communication required to build accurate threat models. Several participants shared the sentiment of P15 who noted: *"We also don't know what other teams are doing."* For instance, several data scientists noted that were not sure how data was gathered or processed: *"In most of the cases I'm not engaged in gathering data"* (P14), and *"I'm not sure whether there are any other security things done at data engineering level"* (P02). Similarly, one data engineer described how they did not have visibility on how the data was used: *"We wouldn't know too much about, what they were going to be developing in terms of like model requirements"* (P13). This separation of ML pipeline parts, and the disconnect of information on how data is collected, processed, and used in the ML model prevents a holistic view of the entire pipeline as one ML system to build a threat model and defenses for. In addition, it prevents each separate team from accurately considering the risks introduced by the other parts of the pipeline in a threat model for themselves. Yet, we find that multiple of our participants considered the data they used as internal, and therefore safe to use, despite having no information if any protective measures were taken: *"We don't do much work on the data side. We just get it from [the client] and then we train"* (P02).

In addition to isolation from other teams in the ML pipeline, we find that several participants were not connected to information security specialists, teams, and departments in their company. For example, P08 described the information security team as *"relatively isolated. We have never worked on a project together, so I would say we live in two different sets of worlds in terms of our focus."* Thus, for several practitioners, S&P teams may not support their implementation of defenses.

#### Key Insights:

- ML practitioners are isolated from other teams in the ML pipeline and S&P teams at their company.
- This isolation may increase the difficulty of both defending against AML attacks that span the ML pipeline and receiving assistance from S&P experts.

**AML Tasks Remain Open Because Responsibility Is Not Defined.** As the responsibility for protecting ML models was undefined or unclear to the participants, AML tasks tended to be overlooked, unassigned, and not implemented.

Overall, participants lacked understanding of what was done and by whom. Regular AML-relevant steps of ML development such as data cleaning were not always well-defined processes. When asked explicitly about responsibility for security, participants said that *"There's no person as such, responsible for the security"* (P02) and *"Protect the ML algorithms to be hacked?...In the companies that I work, nobody cares about this"* (P09). In particular, only one participant reported that there was a dedicated AML team responsible for such concerns.

In the absence of a defined responsible person, there were two primary mindsets. On the one hand, some participants felt a sense of obligation and responsibility of their own accord. P20 reported that they *"went ahead and just removed all of that [private] information, not because anyone asked me to, but just because I thought it should be."* Other participants reported they would *"personally care"* (P03) or had *"a code of honor"* (P06). In contrast, several other participants stuck to what tasks were explicitly assigned to their role and did not concern themselves further: *"I always try to follow the rules that the company gives me"* (P14), and *"I just follow the protocols"* (P11). When confronted with the problem of implementing AML defenses, several of these participants wanted to externalize the responsibility to other roles regardless of these roles' (A)ML knowledge or qualifications, saying *"It's not my problem, actually...That's [the security programmer's] job"* (P17), *"If it creates a bug, it's a web development [responsibility]"* (P01), *"it's already the job of the security guys or the IT guys to maintain most of my models and they deal with these security issues"* (P11), and *"The owner of the company would be responsible for [a breach]"* (P20).

#### Key Insights:

- Responsibility for securing ML models was undefined and externalized, leading to unimplemented defenses.

**AML Defense Competes With Other Business Priorities, and May Lose.** While participants were agreeable to the idea of AML defenses, the implementation of these defenses were found to be in conflict with other business-imposed constraints. Because of this, several participants perceived that *"the priority to defend against adversarial attacks might be lower"* (P16), resulting in vulnerabilities left unmitigated.

Several of our participants were pressured by time constraints during their projects. For these participants, the time needed to implement a defense was viewed as a significant hurdle towards implementation: *“The projects are squeezed into very short time intervals... There is no time nor energy to develop security defenses”* (P17). As noted by P20, it was not only the implementation of such defenses that would be problematic, but the time needed to understand these defenses as well: *“Awareness would be a whole block. And yes, and also time allocated specifically for that.”*

Cost-constraints were also noted by several participants to be a significant barrier towards defense implementation. While hiring AML-specific roles would solve time-constraints for understanding, developing, and implementing defenses, P07 noted that the cost of such experts could be burdensome: *“[You’ll need] three or four very expensive individuals... millions of dollars a year for all the resources. To be honest, I think people would love to have the security, but it’s very expensive to get this kind of know-how to be able to adequately defend against that kind of thing, especially when you’re working against tight budgets.”* Thus, it may be impractical to assume expert AML roles or consulting will be available for smaller, more constrained companies.

As some AML defenses decrease model performance [2, 70, 81], several participants noted that these costs would be a barrier. Occasionally participants noted that accepting degrading performance is conditional on the risk AML posed: *“If there’s such a threat and the project totally collapses, [the customers] would be... ok with accepting the decreasing accuracy”* (P17). However, others noted that that any decreased performance would be unacceptable, particularly for financial-purpose models where losses in accuracy directly translated to losses in revenue: *“[The bank doesn’t] want to lose money... The accuracy is very important for them. So the security guys need to adopt, not the modelers”* (P11). Thus, depending on the application, various teams may be more or less willing to accommodate degraded performance.

#### Key Insights:

- Resource-constraints (such as time and money) restrict the ability for practitioners to implement AML defenses.
- Even if implemented, defenses that harm model effectiveness were considered unacceptable by some practitioners.

**Practitioners Have Difficulties in Finding Applicable Defenses.** When asked how they would implement AML defenses, many participants said they would look for already available, applicable solutions. However, none of our participants knew of or used any such tools. Despite this, several practitioners wished popular ML libraries and platform providers could include defensive implementations for them to import, saying *“I will look for Python packages, if there is already a built-in protecting packages”* (P01), and *“Since I’m using Google Cloud, since I’m deploying my model*

*into Google Cloud, I would first look into the Cloud functions”* (P13). Many participants agreed that having AML defenses available as a part of tools, libraries, and platforms they already use would immensely help them implement protection into their ML systems. This is especially relevant as our participants reported that they often had difficulties using the results of scientific papers for their projects, and thus rely on more productionized, out-of-the box defenses: *“Whenever we try and implement research papers and trying to reproduce their results, it is never straightforward”* (P16). Furthermore, some believed that the inclusion in such libraries could also help bring the topic to the attention of more practitioners: *“There are some tools that are like TensorFlow. They’re very, very popular. So, if they are to call you out for example, mentioned about [AML], then the spread, I think it’s relatively high”* (P10).

#### Key Insights:

- Practitioners have difficulties finding productionized AML defenses.
- Including defenses in popular libraries and platforms may increase their accessibility and exposure.

## 6 Discussion

In our results we identify potential root causes of barriers to AML defense implementations in industry. Key challenges for practitioners begin with poor exposure to AML concepts and effective learning resources (Section 5.1), resulting in misunderstandings of the risks they pose and the vulnerabilities their systems may have (Section 5.2). This is further compounded by practitioners’ lack of awareness around readily available AML tools (Section 5.3), potentially resulting in an inability to evaluate or monitor their systems for AML risks (Section 5.2). Lastly, organizational structures further exacerbated these issues (Section 5.3): isolated between teams resulted in poor understanding of the ML pipeline’s threat model, unassigned responsibility led to a lack of substantive action, and conflicting company objectives demoted the importance of AML mitigations and restricted resources required for defense implementation.

Given the barriers towards defending against this rapidly evolving threat, it is unsurprising that practitioners have not implemented adequate defenses. Furthermore, uncertainties around whether defenses produced today will remain effective in the current AML arms race, may cause additional hesitations and delays. With these constraints in mind, rather than prescribing specific defenses, we believe that it is most pertinent for companies and practitioners to embed the necessary cultural and technical infrastructure to monitor such risks and become prepared to mitigate threats if necessary. Thus, we recommend a set of actions that will increase the awareness of AML risks to stakeholders and preemptively disentangle organizational barriers that might prevent an effective response.



**Establish an S&P Culture in ML.** We find that disconnects between ML and security teams, as well as unclear AML responsibilities are major organizational challenges that prevent a constructive collaboration on AML mitigations (Section 5.3). Similar to prior work in software engineering [8, 97], we notice an over-reliance on other teams to take care of S&P objectives, low consideration of AML security compared to more functional objectives, and a split between S&P-conscious and S&P-avoidant ML developers among participants. Thus, in line with effective recommendations from the software engineering field to overcome such challenges, companies may need to facilitate a shift in culture by promoting advocacy for AML conscientiousness and accounting for such threats in existing lifecycles.

**Introduce and Promote S&P Advocates in ML.** In order to spearhead S&P mindfulness and behaviors, companies should introduce and promote S&P advocates such as informal champions within ML teams or official AML experts to consult with.

“Champions” are generally enthusiastic advocates of a specific software trait such as usability [57], security [61, 85], and privacy [83]. In the field of software engineering, such champions have been proven effective at improving S&P culture by encouraging an awareness of S&P concerns, developer communication with S&P teams, and the adoption of mitigations [83, 85]. Thus, the field of ML may similarly benefit from the promotion of S&P champions. While such champions can be intentionally hired and recruited, simply emphasizing the importance of S&P in ML, providing opportunities for S&P growth, and recognizing, supporting, and rewarding voluntary advocates allow existing developers to grow into such roles [38, 83]. Once found, these champions will not only naturally cultivate an S&P community but can be also intentionally embedded within projects or reviewing teams to provide targeted assistance. This said, while champions may prove helpful for S&P advocacy in ML, they may not necessarily hold the expertise, or resources to implement effective AML defenses. Therefore, we also recommend the introduction of a ML S&P role in organizations. As suggested by participants, having someone to consult with or help implement AML remediations would allow for increased awareness of threats, and importantly, practical implementation of defenses. However, we do acknowledge that such roles are resource-intensive and may not be possible for all organizations to support. In such cases, having someone on the security team with a stronger understanding of ML could still go a long way in identifying and implementing more simplistic AML remediations (e.g., the removal of sensitive training data). Independent of the specific solution, we emphasize the importance of clearly defining and assigning responsibility for AML risks in an organization.

**Integrate S&P Practices into ML Processes.** In addition to advocacy, a commitment to implementing AML-aware procedures can help establish and reinforce an S&P culture in ML.

While several participants maintained processes to evaluate the privacy of used data, and the security of the surrounding software, this did not encompass AML considerations in ML projects. To account for this, Kumar et. al [78] suggests having a set of integrated best practices (or a Secure Development Lifecycle (SDL)) for ML development to identify threats and remediate vulnerabilities. However, our results also suggest that SDLs may improve AML protections in ways other than just threat identification and remediation. First, SDLs may help remediate the de-prioritization of AML noted by participants. By intentionally and proactively allotting time for vulnerability checks, SDLs can allow for a re-prioritization of S&P within traditionally functionality-focused teams [7], a common deterrent for security [8, 86]. Second, SDLs may increase awareness of AML vulnerabilities. Simple acts such as recognizing vulnerabilities from threat modeling procedures or discussions around S&P that may occur in scrum meetings/workshops can increase S&P awareness and allow developers to begin thinking adversarially [86]. Third, SDLs can improve an organization’s S&P. Once vulnerabilities are illuminated and discussions around S&P become more frequent, practitioners may begin viewing functional, but insecure software as of a lower quality, prompting a drive to secure them [86]. However, SDLs must be implemented correctly to ensure continued vigilance. Prior work notes that while S&P can be handled by designated security experts, this reliance may result in an externalization of responsibility [67, 85]. Thus, several works ultimately recommend a collective S&P effort among general developers for any effective long-term security culture [7, 85, 86, 97]. Furthermore, S&P tasks must be explicitly assigned, otherwise, other explicitly assigned priorities may be preferred and lead to a de-prioritization of security over time [86].

**Promote Practitioner AML Awareness.** In order to allow for effective communication and remediation of threats, ML practitioners need to be aware of the potential security and privacy risks in their products. While security experts and champions are helpful in advocating and fixing immediate threats, prior work largely recommends a larger organizational response among the entire developer teams [7, 67, 85, 86, 97]. However, unlike prior work in software engineering, in which a substantial number of developers knew of software vulnerabilities and ways to address them [85], we find that most of our participants did not understand one or more AML concepts prior to our interview (Section 5), likely caused by a lack of exposure in educational and job requirements, and a poor availability of AML resources (Section 5.1).

While it is impractical to require AML expertise of all ML practitioners, even preliminary exposure may yield great benefits. Much like how showing traditional vulnerabilities to software developers encourages a security-oriented mindset [15, 58], exposing ML practitioners to AML concepts may

allow for consideration of AML risks during the ML lifecycle. This exposure can be improved via several vectors.

**Improve Educational Curriculums by requiring coverage of AML concepts.** Drawing on prior works' recommendations to promote a culture of security in software [77, 104] and ethics in AI [30, 71], introducing AML concepts in required ML courses in university and online programs may be an effective way to implement required exposure of otherwise avoided topics. These topics should not only be theoretically-focused but also include practical discussions of S&P risks in ML systems, and how one can model and discover these vulnerabilities in real systems. The focus should not only be on attacks, but show that defenses can be applied to prevent, monitor, and mitigate attacks. Lastly, to maximize the number of practitioners exposed, these topics should be introduced at the undergraduate level/introductory courses rather than keeping them secluded to research or graduate-level courses.

**Expand Educational Resources around AML to better reach practitioners.** Currently, research papers are the primary repository of AML knowledge, but several of our participants reflected on how difficult it is for them to read or implement any ideas found there (Section 5.1). Thus, an effort should be made to expand the available resources for AML learning.

Several participants mentioned hearing about AML solely through blogs and media, yet no participant mentioned continuing to learn about the topic deeper as they did with other ML topics. Furthermore, participants' lack of defense knowledge and the impression that AML threats are futuristic may imply defects in these knowledge vectors. To counteract this, a stronger emphasis should be placed on the practicality of these threats alongside a holistic description of defenses. This may allow for a more complete understanding of AML risks and mitigation to be imparted to practitioners. Furthermore, references to existing, well-supported resources for modeling and managing AML risks [56, 60] should be included to allow for deeper learning of interested readers.

Structured learning via university courses and online courses is also commonly used to learn fundamental ML knowledge, yet participants only reported taking paper-reading courses to learn AML. This may be caused by a lack of such structured resources, since to our knowledge, very few full-length online courses or lecture-based university courses focus on AML. Given the reported effectiveness of courses by our participants and the deeper learning that lecture-based courses encourage compared to student-activated learning [82], a greater effort towards structured, lecture-based AML courses should be made.

Lastly, a variety of resources such as books, community forums, open-source code, documentation, and hackathons were reported to be used for ML learning but not used in AML learning. These gaps represent opportunities for AML education. Future investments in AML education would do

well to begin to provide more such spaces to accommodate a variety of learning styles among practitioners.

**Provide Accessible Monitoring and Assessment Solutions.**

While adoption of defenses is strongly influenced by developer's understanding of their system's vulnerabilities and prior breaches [8, 89], we find that several participants lacked adversarial evaluations to discover vulnerabilities and none of our participants reported monitoring their models for AML exploits occurring on their system (Section 5.2). Furthermore, several participants were resistant to using such tools if they compete with other priorities (Section 5.3). As a result, practitioners, management, and other stakeholders lack the visibility to adequately assess AML risks which inform decisive remedial actions. Thus, we recommend that both academia and industry encourage the adoption of AML risk assessment solutions by making them readily accessible and adoptable to practitioners.

**Increase Toolset awareness.** While several AML toolsets exist [32, 52, 59, 63] none of our participants were aware of any tools. As prior work notes, security tool adoption is heavily influenced by the exposure a security tool has, as well as the trust one has in the source recommending it, with interpersonal/organizational connections being most effective at encouraging adoption [96]. Importantly however, no participant mentioned learning AML-tools from organizational or interpersonal connections. This implies that a large vector of influence is not currently utilized for AML tool adoption and thus, organizational support for defensive toolsets should be increased. This may be accomplished by requiring evaluation of ML models with toolsets as an organizational policy, empowering the security team to suggest AML tools and processes for others, or promoting education of how to use such tools by supporting practitioner attendance in AML workshops/conferences [96]. Furthermore, blog-like mediums were effective vectors for spreading AML knowledge among participants (Section 5.1); however, they did not typically contain deep information or links to actionable defenses, such as toolkits. Similarly, some well-known frameworks for AML threat modeling may include helpful advice, but do not overtly advertise defensive tools [50, 56]. Thus, encouraging direct referencing of available toolkits in commonly read AML media may be a simplistic yet effective step increasing toolkit adoption.

**Solutions must accommodate business-constraints.** Even if toolsets are well-known, functional and monetary costs related to defense usage and implementation may hinder or outright prevent defense deployment (Section 5.3). Thus, tools should attempt to require as little practitioner effort, monetary cost, and functional degradation as possible. Unfortunately, while external monitoring solutions may result in little to no hindrances on a model's effectiveness [21, 45], model and data-intrusive AML defenses that prevent such attacks may directly contend with model utility [88, 101, 105]. Thus, we

currently recommend that companies implement lightweight model monitoring. This will better inform decision-makers of possible exploits and will help guide remedial action and further defense implementations if necessary. Concurrently, researchers and toolset developers should focus efforts to develop defensive solutions with minimal utility costs, so that cost-sensitive companies can protect themselves as soon as possible. Beyond utility, difficulties in implementing and using tools also hinder tool adoption by software developers [94]. Thus, AML tools should be easy to use and easily integrate with existing pipelines to prevent additional burden for developers. This may be readily accomplished by ensuring that tools are well-documented, usable, and ideally rolled out via widely used ML libraries or platforms. Furthermore, as these tools may be used by non-security-oriented practitioners or other non-technical stakeholders, risk reporting should be understandable to non-AML experts and clearly indicate risks. This may be achieved by providing language devoid of complex terminology and including actionable follow-up steps; however, how to clearly communicate AML risks in this way is an unexplored problem. Furthermore, this form of monitoring shouldn't increase other risks. For instance, privacy risks introduced from the monitoring of model queries would need to be carefully considered and resolved.

## 7 Conclusion

Through interviews with 21 ML practitioners, we investigated what barriers impede ML practitioners from implementing effective AML defenses in productionized ML systems. We found that practitioners' ability to deploy defenses is hampered by a lack of institutional motivation and educational resources to learn the concepts, an inability to adequately assess their AML risk and make subsequent decisions, and organizational structures and goals that hinder implementation in favor of other objectives. From these findings, we provided recommendations to meaningfully improve the awareness of ML practitioners to these threats and disentangle organizational barriers that may prevent an effective response to AML threats.

## Acknowledgments

This work is supported in part by the "Responsible Artificial Intelligence in the Digital Society" PhD program funded by the Ministry of Science and Culture of Lower Saxony, Germany as well as the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE – 1746047. The views presented in the paper are those of the authors and do not necessarily reflect the views of any funding agencies.

## References

- [1] Upwork. <https://www.upwork.com/>.
- [2] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proc. of SIGSAC*, 2016.
- [3] Yasemin Acar, Michael Backes, Sascha Fahl, Simson L. Garfinkel, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. Comparing the usability of cryptographic apis. In *Proc. of IEEE Symposium on Security and Privacy*, 2017.
- [4] Yasemin Acar, Michael Backes, Sascha Fahl, Doowon Kim, Michelle L. Mazurek, and Christian Stransky. You get where you're looking for: The impact of information sources on code security. In *Proc. of IEEE Symposium on Security and Privacy*, 2016.
- [5] Ross Anderson. *Security Engineering. A guide to building dependable distributed systems*. Wiley, Hoboken, NJ, 3rd edition, 2020.
- [6] Alexey Antonov and Alexey Kogtenkov. How to confuse antimalware neural networks. adversarial attacks and protection, 2021. <https://securelist.com/how-to-confuse-antimalware-neural-networks-adversarial-attacks-and-protection/102949/>.
- [7] Hala Assal and Sonia Chiasson. Security in the software development lifecycle. In *Proc. of SOUPS*, 2018.
- [8] Hala Assal and Sonia Chiasson. 'think secure from the beginning' a survey with software developers. In *Proc. of CHI*, 2019.
- [9] Giuseppe Ateniese, Giovanni Felici, Luigi V. Mancini, Angelo Spognardi, Antonio Villani, and Domenico Vitali. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *International Journal of Computer and Systems Engineering*, 2015.
- [10] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proc. of ICML*, 2018.
- [11] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *Proc. of ICML*, 2018.
- [12] Christiaan Beek. Virustotal poisoning, 2020. <https://atlas.mitre.org/studies/AML.CS0002>.
- [13] Lukas Bieringer, Kathrin Grosse, Michael Backes, Battista Biggio, and Katharina Krombholz. Industrial practitioners' mental models of adversarial machine learning. In *Proc. of SOUPS*, 2022.
- [14] Zeki Bilgin, Mehmet Akif Ersoy, Elif Ustundag Soykan, Emrah Tomur, Pinar Çomak, and Leyli Karavaş. Vulnerability prediction from source code using machine learning. *IEEE Access*, 2020.
- [15] Matt Bishop. A clinic for "secure" programming. In *Proc. of IEEE Symposium on Security and Privacy*, 2010.
- [16] Igor Buzhinsky, Arseny Nerinovsky, and Stavros Tripakis. Metrics and methods for robustness evaluation of neural networks with generative models. *Machine Learning*, 2021.
- [17] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruiqiang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *Proc. of IEEE Symposium on Security and Privacy*, 2021.
- [18] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proc. of AISec*, 2017.
- [19] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proc. of IEEE Symposium on Security and Privacy*, 2017.
- [20] Jian Chen, Xuxin Zhang, Rui Zhang, Chen Wang, and Ling Liu. Depois: An attack-agnostic defense against data poisoning attacks. *IEEE Transactions on Information Forensics and Security*, 2021.



- [21] Steven Chen, Nicholas Carlini, and David Wagner. Stateful detection of black-box adversarial attacks. In *Proc. of AISec*, 2020.
- [22] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint*, 2017.
- [23] Vanya Cohen, Aaron Gokaslan, Ellie Pavlick, and Stefanie Tellex. Opengpt-2: We replicated gpt-2 because you can too, 2019. [https://medium.com/@vanya\\_cohen/45e34e6d36dc](https://medium.com/@vanya_cohen/45e34e6d36dc).
- [24] Sunny Consolvo, Patrick Gage Kelley, Tara Matthews, Kurt Thomas, Lee Carosi Dunn, and Elie Bursztein. “why wouldn’t someone think of democracy as a target?”: Security practices & challenges of people involved with us political campaigns. In *Proc. of USENIX Security Symposium*, 2021.
- [25] Juliet Corbin and Anselm Strauss. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Sage Publications, Thousand Oaks, CA, 4th edition, 2015.
- [26] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 2018.
- [27] Luca Di Persio and Oleksandr Honchar. Multitask machine learning for financial forecasting. *International Journal of Circuits, Systems and Signal Processing*, 2018.
- [28] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proc. of CVPR*, 2018.
- [29] Michael Feldman, Frida Juldaschewa, and Abraham Bernstein. Data analytics on online labor markets: Opportunities and challenges. *arXiv preprint*, 2017.
- [30] Casey Fiesler, Natalie Garrett, and Nathan Beard. What do we teach when we teach tech ethics? a syllabi analysis. In *Proc. of SIGCSE*, 2020.
- [31] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proc. of ICLR*, 2015.
- [32] Dou Goodman, Hao Xin, Wang Yang, Wu Yuesheng, Xiong Junfeng, and Zhang Huan. Advbox: a toolbox to generate adversarial examples that fool neural networks. *arXiv preprint*, 2020.
- [33] Google. Responsible ai. <https://cloud.google.com/responsible-ai>.
- [34] Kathrin Grosse, Lukas Bieringer, Tarek Richard Besold, Battista Biggio, and Katharina Krombholz. “why do so?”—a practical perspective on machine learning security. In *Proc. of AdvML*, 2022.
- [35] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint*, 2017.
- [36] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *Proc. of ICLR*, 2018.
- [37] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proc. of ICML*, 2018.
- [38] Martin Gilje Jaatun and Daniela Soares Cruzes. Care and feeding of your security champion. In *Proc. of CyberSA*, 2021.
- [39] Kaggle. State of Data Science and Machine Learning 2021, 2021. <https://www.kaggle.com/kaggle-survey-2021>.
- [40] Harjot Kaur, Sabrina Amft, Daniel Votipka, Yasemin Acar, and Sascha Fahl. Where to recruit for security development studies: Comparing six software developer samples. In *Proc. of USENIX Security Symposium*, 2022.
- [41] Mazaher Kianpour and Shao-Fang Wen. Timing on machine learning: State of the art. In *Intelligent Systems and Applications*, 2020.
- [42] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *Proc. of ICLR*, 2017.
- [43] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018.
- [44] Peter Lee. Learning from tay’s introduction - the official microsoft blog, 2016. <https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/>.
- [45] Huiying Li, Shawn Shan, Emily Wenger, Jiayun Zhang, Haitao Zheng, and Ben Y Zhao. Blacklight: Scalable defense for neural networks against Query-Based Black-Box attacks. In *Proc. of USENIX Security Symposium*, 2022.
- [46] Tianshi Li, Elizabeth Louie, Laura Dabbish, and Jason I Hong. How developers talk about personal data and what it means for user privacy: A case study of a developer forum on reddit. In *Proc. of CHI*, 2021.
- [47] Yuanchun Li, Jiayi Hua, Haoyu Wang, Chunyang Chen, and Yunxin Liu. Deeppayload: Black-box backdoor attack on deep learning models through neural payload injection. In *Proc. of ICSE*, 2021.
- [48] Jiajun Lu, Theerasit Issaranon, and David Forsyth. Safetynet: Detecting and rejecting adversarial examples robustly. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.
- [49] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *Proc. of ICLR*, 2018.
- [50] Andrew Marshall, Jugal Parikh, Emre Kiciman, and Ram Shankar Siva Kumar. Threat modeling ai/ml systems and dependencies, 2020. <https://learn.microsoft.com/en-us/security/engineering/threat-modeling-aiml>.
- [51] Nora McDonald, Sarita Schoenebeck, and Andrea Forte. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for and hci practice. In *Proc. of CSCW*, 2019.
- [52] Marco Melis, Ambra Demontis, Maura Pintor, Angelo Sotgiu, and Battista Biggio. secm1: A python library for secure and explainable machine learning. *arXiv preprint*, 2019.
- [53] Microsoft. Responsible ai. <https://www.microsoft.com/en-us/ai/responsible-ai?SilentAuth=1&wa=wsignin1.0&activetab=pivot1:primaryr6>.
- [54] Microsoft. Microsoft - azure service, 2020. <https://atlas.mitre.org/studies/AML.CS0010/>.
- [55] Jaron Mink, Harjot Kaur, Julaine Schmusser, 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 - Replication Package, 2023. <https://osf.io/3q54p/>.
- [56] Mitre. Mitre adversarial threat landscape for artificial-intelligence systems (atlas), 2020. <https://atlas.mitre.org>.
- [57] Deborah Mrazek and Michael Rafeld. Integrating human factors on a large scale: product usability champions. In *Proc. of CHI*, 1992.
- [58] Kara Nance. Teach them when they aren’t looking: Introducing security in cs1. *IEEE Security & Privacy*, 2009.
- [59] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *arXiv preprint*, 2018.
- [60] NIST. AI Risk Management Framework. <https://www.nist.gov/itl/ai-risk-management-framework>.
- [61] OWASP. Owasp security culture, 2022. [https://owasp.org/www-project-security-culture/v10/4-Security\\_Champions/](https://owasp.org/www-project-security-culture/v10/4-Security_Champions/).
- [62] Nicolas Papernot. A marauder’s map of security and privacy in machine learning: an overview of current and future research directions for making machine learning secure and private. In *Proc. of AISec*, 2018.



- [63] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint*, 2018.
- [64] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proc. of Asia CCS*, 2017.
- [65] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *Proc. of IEEE Symposium on Security and Privacy*, 2016.
- [66] Will Pearce and Nick Landers. Proof-pudding, 2020. <https://github.com/moohax/Proof-Pudding>.
- [67] Andreas Poller, Laura Kocksch, Sven Türpe, Felix Anand Epp, and Katharina Kinder-Kurlanda. Can security become a routine? a study of organizational change in an agile software development group. In *Proc. of CSCW*, 2017.
- [68] PwC. Responsible ai toolkit. <https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html>.
- [69] Emilee Rader, Samantha Hautea, and Anjali Munasinghe. "i have a narrow thought process": Constraints on explanations connecting inferences and self-perceptions. In *Proc. of SOUPS*, 2020.
- [70] Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Proc. of NeurIPS*, 2020.
- [71] Jeffrey Saltz, Michael Skirpan, Casey Fiesler, Micha Gorelick, Tom Yeh, Robert Heckman, Neil Dewar, and Nathan Beard. Integrating ethics within machine learning courses. *ACM Transactions on Computer Education*, 2019.
- [72] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *Proc. of CHI*, 2021.
- [73] Adi Shamir, Itay Safran, Eyal Ronen, and Orr Dunkelman. A simple explanation for the existence of adversarial examples with small hamming distance. *arXiv preprint*, 2019.
- [74] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proc. of CCS*, 2016.
- [75] Reza Shokri, Marco Stronati, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proc. of CCS*, 2017.
- [76] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert D. Mullins, and Ross J. Anderson. Sponge examples: Energy-latency attacks on neural networks. In *Proc. of EuroS&P*, 2021.
- [77] Ambareen Siraj, Sheikh Ghafoor, Joshua Tower, and Ada Haynes. Empowering faculty to embed security topics into computer science courses. In *Proc. of ITiCSE*, 2014.
- [78] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissioneru, Matt Swann, and Sharon Xia. Adversarial machine learning-industry perspectives. In *Proc. of IEEE SP Workshops*, 2020.
- [79] Micah J. Smith, Roy Wedge, and Kalyan Veeramachaneni. Featurehub: Towards collaborative data science. In *Proc. of DSAA*, 2017.
- [80] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *Proc. of ICLR*, 2018.
- [81] Jacob Steinhardt, Pang Wei Koh, and Percy Liang. Certified defenses for data poisoning attacks. In *Proc. of NeurIPS*, 2017.
- [82] Katrien Struyven, Filip Dochy, Steven Janssens, and Sarah Gielen. On the dynamics of students' approaches to learning: The effects of the teaching/learning environment. *Learning and instruction*, 2006.
- [83] Mohammad Tahaei, Alisa Frik, and Kami Vaniea. Privacy champions in software teams: Understanding their motivations, strategies, and challenges. In *Proc. of CHI*, 2021.
- [84] MITRE AI Red Team. Mitre - physical adversarial attack on face identification, 2020. <https://atlas.mitre.org/studies/AML.CS0012>.
- [85] Tyler W. Thomas, Madiha Tabassum, Bill Chu, and Heather Lipford. Security during application development: An application security expert perspective. In *Proc. of CHI*, 2018.
- [86] Kerry-Lynn Thomson, Rossouw von Solms, and Lynette Louw. Cultivating an organizational information security culture. *Computer Fraud & Security*, 2006.
- [87] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *Proc. of USENIX Security Symposium*, 2016.
- [88] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *Proc. of ICLR*, 2019.
- [89] Sven Türpe, Laura Kocksch, and Andreas Poller. Penetration tests a turning point in security practices? organizational challenges and implications in a software development team. In *In Proc. WSIW*, 2016.
- [90] Daniel Votipka, Desiree Abrokwa, and Michelle L. Mazurek. *Building and Validating a Scale for Secure Software Development Self-Efficacy*. 2020.
- [91] Eric Wallace, Mitchell Stern, and Dawn Song. Imitation attacks and defenses for black-box machine translation systems. In *Proc. of EMNLP*, 2020.
- [92] Charles Weir, Ben Hermann, and Sascha Fahl. From needs to actions to secure apps? the effect of requirements and developer practices on app security. In *Proc. of USENIX Security Symposium*, 2020.
- [93] Zack Whittaker. Security lapse exposed clearview ai source code, 2020. <https://techcrunch.com/2020/04/16/clearview-source-code-lapse/amp/>.
- [94] Jim Witschey, Shundan Xiao, and Emerson Murphy-Hill. Technical and personal factors influencing developers' adoption of security tools. In *Proceedings of the 2014 ACM Workshop on Security Information Workers*, 2014.
- [95] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proc. of ICML*, 2018.
- [96] Shundan Xiao, Jim Witschey, and Emerson Murphy-Hill. Social influences on secure development tool adoption: why security tools spread. In *Proc. of CSCW*, 2014.
- [97] Jing Xie, Heather Richter Lipford, and Bill Chu. Why do programmers make security errors? In *Proc. of VL/HCC*. IEEE, 2011.
- [98] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Proc of NDSS*, 2018.
- [99] Weilin Xu, Yanjun Qi, and David Evans. Automatically evading classifiers: A case study on PDF malware classifiers. In *Proc of NDSS*, 2016.
- [100] Henry Xuef et al. Camera hijack attack on facial recognition system, 2020. <https://atlas.mitre.org/studies/AML.CS0004>.
- [101] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *Proc. of NeurIPS*, 2020.

- [102] Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *Proc. of ICML*, 2019.
- [103] Bin Yu, Jie Pan, Jiaming Hu, Anderson Nascimento, and Martine De Cock. Character level based detection of dga domain names. In *Proc. of IJCNN*. IEEE, 2018.
- [104] Chuan Yue. Teaching computer science with cybersecurity education built-in. In *Proc. of ASE*, 2016.
- [105] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proc. of ICML*, 2019.

## A Screening Questions

"DS" denotes the questions asked to data scientists and "DE" denotes the questions asked to data engineers.

- 1.A. (DS) Do you have experience in predictive data science or machine learning within a company setting? If so, how many years?
- 1.B. (DE) Do you have experience in data collection or processing for predictive data science or machine learning within a company setting? If so, how many years?
2. (DS + DE) Can you briefly describe your experience? (two sentences is enough)

## B Participant and Demographics

Number of Participants	21
<b>Gender:</b>	
Female	4
Male	17
<b>Age:</b>	
Mean	27.3
std	4.6
<b>Employment Status:</b>	
Employed full-time	12
Employed part-time	1
Independent contractor, freelancer, or self-employed	3
Not employed, but looking for work	2
Student	3
<b>Education:</b>	
Bachelor's degree (B.A., B.S., B.Eng., etc.)	11
Master's degree (M.A., M.S., M.Eng., MBA, etc.)	8
Other doctoral degrees (Ph.D., Ed.D., etc.)	2
<b>Degree Field*†:</b>	
Electrical/Computer Engineering	5
Computer Science	3
Data Science	2
AI	1
Statistics	1
Systems	1
Information Systems	1
Aerospace Engineering	1
Mechanical Engineering	1
Mechatronic Engineering	1
Biology	1
Finance	1
Math	1
Physics	1
<b>Location:</b>	
USA	4
India	3
Turkey	2
Austria	2
Australia	1
Philippines	1
Netherlands	1
Pakistan	1
New Zealand	1
Canada	1
UK / N. Ireland	1
Poland	1
Algeria	1
South Africa	1
<b>Ethnicity*:</b>	
White or of European descent	9
South Asian	5
Hispanic or Latino/a/x	1
Middle Eastern	5
East Asian	2
Southeast Asian	1

Table 2: **Participant demographics** Detailed demographics of our participants (\*multiple answers allowed, †open-ended answers)