Towards More Practical Threat Models in Artificial Intelligence Security

Kathrin Grosse,[1] Lukas Bieringer,[2] Tarek R. Besold,[3] Alexandre Alahi[1]
[1]*EPFL, Switzerland,* [2]*QuantPi, Germany,* [3]*TU Eindhoven, The Netherlands*

Abstract

Recent works have identified a gap between research and practice in artificial intelligence security: threats studied in academia do not always reflect the practical use and security risks of AI. For example, while models are often studied in isolation, they form part of larger ML pipelines in practice. Recent works also brought forward that adversarial manipulations introduced by academic attacks are impractical. We take a first step towards describing the full extent of this disparity. To this end, we revisit the threat models of the six most studied attacks in AI security research and match them to AI usage in practice via a survey with **271** industrial practitioners. On the one hand, we find that all existing threat models are indeed applicable. On the other hand, there are significant mismatches: research is often too generous with the attacker, assuming access to information not frequently available in real-world settings. Our paper is thus a call for action to study more practical threat models in artificial intelligence security.

1 Introduction

A large body of academic work focuses on machine learning (ML) security [5, 9, 14, 17, 19, 32, 40, 55, 57, 68, 71]. Although these attacks have been established, increasing criticism targets their threat models. For example, most academic papers focus on standalone models [14, 19, 32, 40, 55, 57, 68, 71], while models in practice are generally embedded into pipelines or larger systems [7, 24]. In addition, it has been pointed out that attacks in practice do currently not require the degree of complexity inherent to academic publications [1, 30]. Also, the measurement of manipulations introduced by an attacker was deemed impractical [1, 27], and the overall amount of data available to the attacker in some cases [17, 30].

For example, poisoning attacks [8, 17, 61] require manipulating the training data. Grosse et al. [30] reported cases of poisoning in the wild—yet it is unknown which fraction of companies allow access to their training data. Thus, the number of organizations vulnerable to poisoning attacks is, in practice, unknown. In addition, companies may only allow access to a fraction of their data—another limiting factor for an attack to succeed. As an example, consider a company where 1% of the data can be accessed by the attacker. Most academic attacks require access to more data [17], limiting their usefulness. Analogously, evasion attacks were reported in the wild [30]. Evasion requires the submission of at least one perturbed test sample [19, 68]. Yet the number of AI systems in practice where this is possible is again unknown.

These works illustrate mismatches and demonstrate that some aspects of threat models are unaligned between research and practice. The underlying problem, an absence of knowledge on how artificial intelligence (AI) is used in practice, is however still unaddressed. In other words, it remains unknown whether researched threat models are *representative* of AI usage in practice. We thus take a first step towards measuring this mismatch of AI security research and practice.

Contributions. To this end, we describe the commonly used academic threat models of the six most studied attacks in AI security: poisoning [17], backdoors [17], evasion or adversarial examples [19, 68], model stealing [71], membership inference [15, 65], and property inference [4, 39] in Section 2. To measure whether threat models match practical usage, we design a questionnaire in Section 3 that collects information relevant to AI security like access patterns, data sources, etc. In the same section, we present our sample of **271** AI practitioners, before analyzing our results in Section 4.

We summarize our key findings in Table 1. First, all six analyzed attacks are relevant in practice. In our sample, access to training data and the model is often constrained in practice, indicating overly generous assumptions in researched threat models. This includes large fractions of accessible training data for poisoning and backdoor attacks, and large query budgets for black-box evasion and model stealing. Since there are attacks with low budgets, vulnerabilities can be exploited in practice and mitigations are needed. Other mismatches between practice and research concern the used data, where academic datasets cover a large part of industrial datasets, but some cases are rarely studied. Finally, we aim to understand

Table 1: Key Findings of our work.

	Key Finding (KF)	Section
KF 1	**Access** is generally given to query the model and the corresponding outputs, or not at all.	4.1
KF 2	**The underlying assumptions of all six attacks studied are relevant in practice.**	4.2
KF 3	Scientific threat models tend to be **too generous**:	4.2
KF 3.1	**Poisoning** and **backdoor** threat models assume unpractical fractions of alterable training data.	4.2.1
KF 3.2	**Black-box evasion** and **model stealing** threat models assume unpractical amounts of queries.	4.2.2
KF 3.3	**Model stealing** and **privacy attacks** threat models' assumptions do not represent practical AI usage.	4.2.2
KF 4	**Datasets** have often fewer features in practice than in AI security research.	4.3.1
KF 5	**Code libraries** used in AI are security relevant.	4.3.2
KF 6	**AI security knowledge** does not influence the practical threat models of AI in our sample.	4.3.3

which factors influence threat models in practice. Here, knowledge of AI security has no influence. Only AI knowledge and AI maturity of the company are negatively correlated with public data sources, showing the need for future work.

We then revisit the limitations of our approach (Sect. 5). For example, our paper only provides initial insights about vulnerabilities via public access. Real-world threat surfaces may be larger. Still, the implications of our study (Sect. 6) go beyond the above-discussed shortcomings of threat models. Implications also relate to current legislative attempts like the EU AI Act that requires security and vulnerability assessments of AI systems. We also set previously low numbers of AI security incidents into context and pave the way toward a deep understanding of what affects the security of AI-based products in practice. We then review related work (Sect. 7) and conclude our contributions (Sect. 8).

Remark. *This work should not be interpreted as a finger-pointing exercise. So far, AI security research has relied on best practices of security threat modeling, and we confirm that all 6 studied settings are applicable in practice. However, we describe unstudied settings hoping that we, as a community, can progress together toward more practical research.*

2 Background

Before we review AI threat models, we define AI. To this end, we use the example of machine learning (ML), a subdiscipline of AI, and then outline the differences to other paradigms like reinforcement learning (RL) or data mining (DM). A typical task in machine learning is image recognition, e.g., classifying images from cats and dogs. In this case, we have a dataset of images X and corresponding labels Y with the individual image x and label y. On this data, we train a classifier F defined by its weights ω. We adjust these weights ω during training so that $F(\omega, X) \approx Y$. The classifier F then generalizes to unseen test images $x \in X_t$ and correctly predicts their labels $y \in Y_t$. In contrast to the concrete label output used in this example, y does not have to be discrete ('cat', 'dog') but can be continuous (regression) or more complex (in object detection or image segmentation). In the following, we refer to training data or X, Y for any data used for model training. Test data, or X_t, refers to input during deployment (e.g. after the model development is complete). Test outputs, or Y_t, to the corresponding outputs for a given X_t.

RL, in contrast to ML, learns a policy that determines the behavior of an agent in an environment. Albeit different, also RL requires training and test data which can however take the form of an environment generating this data. DM analyses data and does not necessarily rely on test data. The definitions of deployment data X_t or training data X above are thus broad and encompass different formats like confidence scores and top-one outputs, as well as possible pre-processing. We skipped these details to encompass different paradigms and leave a detailed study of these aspects for future work.

Before we review existing attacks on AI, we describe the existing ML threat model commonly used for these attacks.

2.1 Threat Modelling Artificial Intelligence

In general, we distinguish three different aspects defining an attacker's behavior, its *knowledge*, *capabilities*, and *goal* [9]. We summarize these aspects of AI attacks from academic literature in Table 2, and first review the properties before we discuss different attacks in the following subsection.

Knowledge. This aspect describes what the attacker *has access to* or *has knowledge about*. The training (X,Y) or test (X_t,Y_t) data are examples of information the attacker might have. In some cases, knowing this data roughly may suffice: when the victim is training an image classifier, some images can be sufficient to mount an attack, but not the same images the victim trained on are required. Independent of the data, the attacker may know the model's parameters (ω).

Capabilities. In contrast to knowing or observing system properties, threat modeling also describes what an attacker can *alter*. The attacker may, for example, change input samples at training (x) or test time (x_t) or both. In case we are dealing with ML or RL, it might be relevant to distinguish samples (x) and labels (y), e.g., inputs and associated classes or desired behaviors. Lastly, the attacker may feed the model inputs and

Table 2: Threat models for AI security. Below, we list the attacker's knowledge, capabilities, and goals. For each attack, we denote which knowledge in terms of training data (X,Y), test data (X_t,Y_t), parameters (ω), and classifier's outputs ($F(\omega,x)$) are required. Concerning capabilities, we denote whether the attacker can alter training (x) or test (x_t) samples, labels of samples (y), or observe the output of the model ($F(\omega,x)$). For all properties, we denote required (●), sometimes required (◐), and not required (○). We then denote with ✓whether the goal of the attack is availability (Av.), integrity (Int.), or confidentiality (Conf.).

	Knowledge			Capabilities				Attacker's goal			
	X,Y	X_t,Y_t	ω	x	y	x_t	$F(\omega,x)$	Av.	Int.	Conf.	Description
Poisoning, bilevel [17]	●	○	◐	●	◐	○	○	✓			Decrease performance
Poisoning, label flip [17]	●	○	◐	○	●	○	○	✓			Decrease performance
Backdoor [17]	●	●	◐	●	◐	●	○		✓		Misclassify samples with trigger
Evasion, white-box [68]	○	●	●	○	○	●	○		✓		Misclassify perturbed sample
Evasion, black-box [50]	○	●	○	○	○	●	◐		✓		Misclassify perturbed sample
Model Stealing [56]	○	◐	○	○	○	◐	●			✓	Copy model without consent
Mem. Inf. [39]	○	●	○	○	○	◐	●			✓	Infer sample membership
Attribute Inf. [39]	○	○	●	○	○	○	○			✓	Infer training data attributes

observe the corresponding outputs ($F(\omega,x)$).

Attacker's goal. There are three principal goals [9]. Harming *availability* decreases overall performance to a degree where this system may not be usable anymore. Targeting *integrity* preserves the original performance, but specific outputs may be processed incorrectly (e.g., misclassified). The third, *confidentiality*, concerns the intellectual property of the model and the secrecy of the training data.

Practical concerns - 3rd parties. In research, knowledge and capabilities are often binary (present/not present). In practice, there may be different access levels. In our questionnaire, we consider two levels of accessibility: On the one hand within a company, encompassing employees and clients; and on the other hand 3rd party, or anyone.

Finally, cost-driven assessments should be an additional dimension in analyzing ML security in practice [1]. As both attackers and defenders operate with a cost/benefit mindset [72], attacks will only be conducted if their benefit exceeds the costs. Similarly, defenses will only be applied if their implementation costs are lower than the respective attack's monetary impact that materializes with a certain likelihood [38].

2.2 AI Security

Most AI security work focuses on ML. We thus introduce the ML-security threat models and then discuss the same attacks on other paradigms like RL and DM. We start with training time attacks like poisoning and backdoors and then discuss test time attacks like evasion, and attacks breaching confidentiality like model stealing, membership inference, and data extraction. We focus on the attacks of AI security that received the most attention and visualize them in Table 2.

Poisoning. In poisoning, the attacker alters training data [61] or labels [8] to decrease accuracy, thus targeting availability. Attacks that uniquely target labels are called label-flip attacks [8], whereas poisoning based on the bilevel formulation alters only samples or samples and labels [17]. Alternatively, in sloth attacks, the goal is to increase the model's runtime [16]. Defending poisoning is well understood [17]. Poisoning attacks [6] and defenses [59] have been studied on RL, DM [52], clustering [76], principal component analysis [61], or feature selection algorithms [76].

Backdoors. An alternative attack during training time are backdoors. Backdoors are chosen input patterns that reliably trigger a specified classification output, harming integrity. There are several ways to introduce backdoors [17], via the training [26] or the fine-tuning data [63]. Alternatively, a backdoored model can be provided [22]. Mitigating backdoors has led to an arms race [69], where proposed defenses are broken, leading to new, stronger attacks which again have to be mitigated [17]. Backdoors have also been studied on RL [43].

Evasion/adversarial examples. Evasion decreases the test-time accuracy of a trained and otherwise well-performing classifier [19, 68], and thus also target integrity. To this end, the attacker needs access to the test data and knowledge about the model for white-box attacks, as visualized in Table 2. An exception are black-box attacks, which only require access to the model outputs and knowledge about the rough nature of the data [50]. Alternatively, an attack can be computed on one model F_1 and then transferred to a second classifier F_2 to which the attacker does not have access [57]. Recent works emphasize the need to correctly evaluate defenses [18, 70]. Evasion has also been introduced [46] and tentatively defended [51] on RL and on clustering algorithms [41].

Model stealing. In model stealing, the attacker has black-box access to an ML model and copies its functionality without consent of the model's owner [71] and thus harms confidentiality, as visualized in Table 2. Most model stealing attacks require submitting specific test queries [56], and only one paper [57] obtained models by labeling data from the task the model was purposely used for. In general [56], model stealing attacks are measured by the number of queries they

need and how faithful they reproduce the original model. Similar to model stealing attacks is model extraction, where specially crafted inputs allow the attacker to deduce architectural choices like the usage of dropout [39,55]. Analogous to previous attacks, defenses have been proposed against both attacks, but are caught in an ongoing arms race [56]. RL models can also be stolen [13].

Membership inference. The following attacks target the privacy of the used training data at test time [35, 39]. For example, membership inference [15, 65] predicts membership to the training data for an existing sample based on the target model's output. To this end, attacks rely on membership metrics [65] or shadow-models [65] trained on known membership outputs. Alternatively, repeatedly querying the victim is possible [15]. For all of these attacks, defenses have been proposed [39]. Membership inference has also been demonstrated to work on RL [29]. Beyond membership, inversion attacks attempt to regenerate the training data based on a generative model trained with the victim's outputs [77]. Intuitively, training the model encompasses a large amount of labeled data, which may differ from the original training data.

Attribute inference. In contrast, in attribute inference, the attacker is interested in a specific sensitive attribute or feature. These attacks are mounted assuming white-box knowledge of the victim and using the weights for a meta-classifier [4]. For these attacks, defenses have been proposed [39], but to the best of our knowledge, no works study attacks on RL or DM.

3 Methodology

Having gained an overview of the existing threat models in AI security research, we can now design a questionnaire and decide on a target group to recruit from to assess practical AI threat models. In this section, we first describe the questionnaire design and content, the pretests, and the recruiting procedure, and conclude with the sample description.

3.1 Measuring Threat Models in Practice

In the previous section, we discussed the attacker's knowledge, capabilities, and goals. To assess threat models in practice, we consider the threat model in Table 2 and determine whether knowledge and capabilities are reasonable assumptions in practice. For example Poisoning [17] and backdoor attacks [17] require access and knowledge of training data. Backdoors additionally require the submission of test data to exploit the backdoor. To validate these threat model requirements, we ask industrial practitioners whether a 3rd party can access training inputs and deployment inputs. For backdoors, we ask additional questions about model re-use, a common assumption [22, 63]. For evasion [19, 50, 68], knowledge and access to test data are necessary, while access to the model is optional. We thus also inquire about access to the model from our participants. This access is also implicitly relevant to an attack like model stealing [56, 71]. If access to the model is possible, the attack is superseded. To conclude, the combinations of access control responses help us to determine whether an attack is possible (and, for model stealing, is necessary).

3.2 Questionnaire Design

In other words, we focused on questions concerning the accessibility of the model and data, as these are essential components of threat models (Sect. 2.1). Furthermore, previous work indicated access to models and data as a limiting factor [30]. As questions can be sensitive, we opted for an anonymous survey with 43 questions. For fast completion, the questionnaire only contains multiple choice questions, checkboxes, and relevance rankings based on a Likert scale. Questions, descriptions, and the wording of answer options for multiple-choice questions were based on prior research. In the following, we detail references used for the questionnaire along its three parts, (1) demographics, (2) AI projects, and (3) AI security. The complete questionnaire can be found in the Appendix.

Demographics. We inquired the necessary data to compare to previous studies [30] and populations [42], including gender, age, educational background, company size, AI experience, industry areas[1] and team size [62]. We inquired about our participants' location based on dial codes to obtain privacy-preserving groups. These groups consisted of North, Central, and South America, North/Central, South, and East Europe, Africa, North, East, South/Central, and West Asia (with the Arabian Peninsula), and Australia and Oceania.

AI projects. In this part, we asked questions about threat models like as access to model components and sizes of the used data. We also inquired about other specifications such as the need for a domain expert, time constraints of the application, other specification of requirements for ML model [54, 62], and the possibilities to enforce constraints on the training data [62]. In case the participants worked with several AI-based projects, we asked them to here focus on one.

AI security. This final part focused on the relevance of AI security, privacy, and a self-estimated likelihood of noticing an attack. We also asked whether participants had encountered an attack and what the attack consisted of, as Grosse et al. [30].

3.3 Pretests and Recruiting

After obtaining permission from our institution's ethical review board, we performed two rounds of pretests. All pretesters had AI industry experience (including ML and RL, for example) and were from the author's private networks. The testers were given the questionnaire and asked to think out loud while filling it, enabling us to spot misunderstandings and unclarities. In the first two rounds, 6 participants

[1] https://en.wikipedia.org/wiki/Economy_of_the_United_States_by_sector

(one female, five male) took part. We received minor comments; all questions except three were well understood. We improved these and retested them with four fresh testers. After incorporating the minor changes their feedback agreed on, we implemented the survey in RedCap [34] and started to recruit.

We advertised the study on social media channels such as Twitter and LinkedIn and initially reached out to personal contacts. We then followed previous studies approaches [30] and recruited within AI Slack communities (MLOps, MLSecOps, Pyladies) or contacted potential participants via LinkedIn. We did not impose specific selection criteria other than currently working with AI and did not share the questionnaire if potential participants stated to work only with, for example, ChatGPT. In other words, we targeted practitioners directly involved with AI/ML models or data engineering. While our conclusions affect AI security, the questionnaire is independent of security questions, allowing us to draw from a broader population than previous studies [7, 30]. This was confirmed as we received no feedback that the questions within the questionnaire were unknown to the participants, although we did not screen for a security background. We opted against paying the participants to avoid money-driven participation. Still, many participants were eager to contribute due to their interest in the topic. Throughout the recruiting process, we monitored the gender ratio to ensure the sample remained representative.

We recruited for two and a half months[2] and allowed inputs for one more week to allow potential latecomers to participate. In total, 271 participants filled out our survey.

3.4 Sample Description

A total of 271 participants filled out our questionnaire, of which 201 replied to all questions, and 70 submitted only part of the questionnaire. We do not exclude participants with partial replies. Instead, we report the fraction of participants not providing a reply for the question(s) discussed. Before we analyze the results, we describe the individual and organizational backgrounds of our participants and establish that our sample matches the larger population of AI practitioners.

Individual background of participants. Of our 271 participants, 76% were male, 18.1% female, and the remainder did not reply or did not disclose their gender. Albeit the sample is largely male, the ratio is comparable to similar studies [30] and representative of the population of AI practitioners [42].

The distribution of participants' age was primarily between 25 and 44, with most being between 25 and 34 (44.3%). As before, this distribution matches similar studies [30, 42]. To maintain anonymity, we asked for our participants' locations based on dial codes grouped into twelve areas. We received at least one participant from each area, our sample thus covers the entire globe. Most participants were from Southern (19.9%) and Northern Europe (28%) and North America (18.8%). The fewest participants were from Central America (0.4%), Russia/Mongolia (0.7%), and South America (1.1%). 7.4% did not provide a location. The distribution of academic degrees, with the largest group of master degrees (46.5%) roughly mirrors previous distributions [30, 42]. In terms of AI background, 5.2% were trained only, with most participants (37.3%) having 2-5 years of working experience in AI or ML. Almost as many (35.8%) worked for more than 5 years. Intriguingly, our distribution matches more closely the US-focused distribution than the global distribution of prior work [42], possibly showing a bias of our sample towards Western countries. In terms of team size, most of our participants worked in teams of 6-9 (27.3%) or 3-5 (25.5%) people, less in small teams (<3, 17%) or in teams of 10-15 (12.9%) or even larger than 15 people (14%). This contrasts previous studies [42], which report a quarter of their population in either very small or very large teams.

Organizational background of participants. Although three quarters (77.1%) of our participants' companies were headquartered in North America or Europe, our sample also encompassed companies from Africa (2.2%), Latin America (0.4%), North (0.7%), West (3%), South (5.5%) and East Asia (2.5%) and Oceania (2.2%). Of these companies, roughly every tenth (9.2%) was in automotive or a supplier of automotive, about every eighth in cybersecurity (13.7%), and roughly every seventh in healthcare (15.5%). Other areas encompassed education (3.3%), arts and entertainment (3,3%), and finance and insurance (4,8%). The remainder were other areas. Concerning company size, most participants were from small companies (<50 employees, 34%). Second most were employed at large companies (>1,000 employees, 28.4%), the remainder were in between, coherent with previous studies [30, 42]. AI maturity also coincided with previous samples [30,42]: Few (4.4%) participants stated to work indirectly with AI, most (51.7%) had models in production. Significantly fewer (17.7%) were getting models into production, starting development (11.3%), or evaluating use cases (7%).

4 Results

Knowing that our sample matches the underlying population, we analyze the information our participants provided. We start with the overall threat surface, match the individual attacks, and then report findings beyond specific attacks that help to reconcile security research and practical usage of AI.

Overall threat surface. We start with an overview of common access patterns to model and data as reported by our participants. The most frequent is to allow querying the model and obtaining the corresponding results, or not to give any access to either data, model, or outputs.

Attack specific threat models. We then investigate the six attacks explained in the Background section. As some share significant parts of their threat models, we discuss them to-

[2] Recruiting period: April, 21st to July, 6th, 2023.

Table 3: Most frequent knowledge or capabilities available within our sample. For the components training (X,Y) and test queries (X_t,Y_t), model weights (ω), and model outputs ($F(\omega,x)$), we depict access for 3rd party (●), and no access (○). 25.1% of replies did not cover all questions and are not included, rare combinations with <1.5% are not listed either.

	X,Y	X_t,Y_t	ω	$F(\omega,x)$
32.8%	○	●	○	●
25.1%	○	○	○	○
7.4%	○	●	○	○
2.6%	●	●	○	●
1.8%	○	●	●	●
1.8%	●	●	●	●

gether. For example, poisoning and backdoors both occur during training (Sect. 4.2.1). All other attacks take place at test time (Sect. 4.2.2), where membership inference and attribute inference are grouped as they are both related to privacy (Sect. 4.2.3). For each attack, we compare if the scientific threat model matched our sample's statistics. This was always the case, although there were also significant mismatches.

AI security beyond specific attacks.. Finally, we analyze specific details affecting AI security: dataset sizes, involvement of domain experts, real-time requirements (Sect. 4.3.1), library usage (Sect. 4.3.2), and factors related to the access to AI systems (Sect. 4.3.3). As for the individual attacks, we find both alignments as well as significant mismatches.

4.1 Overall Threat Surface

Before the individual attacks, we describe the overall threat surface using frequent access patterns within our sample.

Threat surface. We compared the threat models commonly used in research, as described in Table 2 in Section 2.1, with the replies of our participants in Table 3. In this table, we reported access in practice to training data (X,Y), test data to query (X_t,Y_t), the model's parameters (ω), and the classifier's outputs ($F(\omega,x)$) ($Q23$,$Q24$,$Q32$,$Q33$). We are interested in *frequent combinations* of allowed access. 68 participants, or 25.1%, did not provide a reply in at least one field. The remainder replied to all four questions. Almost a third of the participants (32.8%) gave access to test queries and model outputs. The second largest combination, with 25.1%, gave no access to data, queries, models, and outputs. The third largest group with 7.4% allowed queries only. Smaller groups contained diverse combinations, including access to everything except the model (2.6%), all but the training data (1.8%), or everything (1.8%). Rare combinations included two cases where nothing but the training data was accessible, and one case with all available except outputs. We now discuss in more detail the individual attacks with their threat surface.

Take away–Overall threat surface. The most frequent access patterns are access to queries and query outputs (32.8%) or no access at all (25.1%).

4.2 Attack Specific Threat Models

Having described the overall practical attack surface, we now focus on the individual attacks' threat models. We study the six attacks described in Sect. 2.2, which we regroup to take into account threat model similarities. We start with training time attacks and review both poisoning and backdoor attacks. Afterwards, we focus on test-time attacks like evasion and model stealing and finally discuss privacy breaching attacks like membership and attribute inference.

4.2.1 Training-Time Attacks

Both poisoning and backdoor attacks perturb the training data to affect the resulting model (compare Table 2). Consequently, the question is how often the training data is accessible ($Q23$). Of our participants, 71.6% reported that the training data was not accessible, and 6.6% that the data was publicly accessible.

These numbers reflect access to the final training data—it might still be possible to tamper with the data at its public origin; when data comes for example from the internet. To this end, we investigated combinations of inaccessible training data ($Q23$) and the percentage of training data from public sources ($Q28$). Here, 100% corresponds to the subset of all participants who reported that their training data was not accessible. The largest group (47.1%) kept their data inaccessible and did not use any data from public sources. Yet, 6.6% stated that 1%-5% of their training data came from public sources. The same held for 5%-10% (9.1%), 10%-15% (4.1%), and 25%-50% (5%) training data from public sources (of our participants). Also, higher percentages like 50%-75% (7.4%) or higher than 75% (10%) of the training data were from public sources even if the resulting data was inaccessible, outlining the need for a complex consideration of practical data security risks.

On the other hand, only 18% of our participants reported that more than 50% or an unknown amount of the data stemmed from public sources. This may indicate that from a practical point of view, relying on high percentages of clean data for defenses is possible. Yet, data quality may then be a problem, and this may be a poor security design choice.

Poisoning. Cinà et al. [17] surveyed the percentage of training data an attacker altered in poisoning attacks. Albeit their analysis focused on vision tasks, their overview summarized common attack assumptions in poisoning. Of their 16 analyzed poisoning papers, the majority (13) tampered with 10-30% training data. The remaining 3 papers altered even more data. We compared these numbers to the percentage of training data from public sources ($Q28$), knowing the amount of

Table 4: Comparing assumptions about alterable training data in poisoning [17] and backdoors [17] to our sample. We state the percent of training data that can be altered, the amount of poisoning and backdoor papers with this specific assumption. Finally, we show the percentage of participants in our sample with this amount of alterable data. The percentages marked with * were misaligned with our questionnaire and were thus estimated. There were 20.7% missing replies in this question.

Percent training data altered	# Poisoning papers [12, 17]	# Backdoor papers [17]	Our findings
>30%	3	—	* <30.3%
10-30%	13	20	* <10.7%
<10%	1	12	20.4%
∅	—	—	30.3%

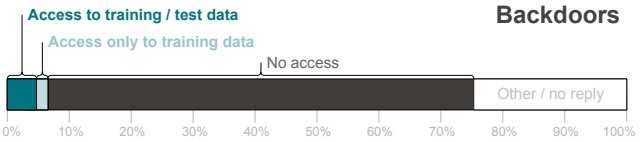

Figure 1: Backdoor threat model in percent of our participants' replies. We report 3rd party access: White denotes incomplete data or an irrelevant threat model (e.g., only test data accessible). Black represents no access, turquoise the backdoor threat model. Light turquoise denotes insufficient access for backdoors, but sufficient access for poisoning attacks.

accessible training data was low. In our sample, either a large fraction of the data came from public sources or none. This contrasts with the existing poisoning threat models, where most papers studied a setting altering 10-30% training data. Many poisoning papers thus studied settings that were rare in practice according to our sample.

Backdoors. Analogous to poisoning attacks, the same survey [17] also covered backdoor attacks. Of 32 systematized papers, about two-thirds (20) tampered with 10-30% of the training data. While no paper altered more data, the remaining 12 papers perturbed less than 10% data. As before, we compared these results to the percentage of training data from public sources ($Q28$). As before, the heavily studied middle range (10-30%) was the least common in practice.

To exploit the backdoor, the attacker must access the test data. We thus investigated combinations of training and test data access within our sample and visualized the results in Figure 1. Of our participants, 6.6% reported training data was accessible to a 3rd party. However, adding the constraint of accessible test data, this reduced to 4.7%; a low attack surface towards backdoors. The setting where only training data is available is studied in poisoning or triggerless backdoors, which instead target a small group of clean samples [17, 26]. Of 32 papers, 10 rely on this specific threat model [17].

Table 5: Comparing assumptions about queries in black-box evasion [50] and model stealing [56] papers. We state the number of queries that can be submitted, then the amount of black-box evasion and model stealing papers assuming the specific amount. Finally, we state our participants reported query amounts, with 19.1% missing replies.

Possible queries	# Evasion black-box papers [50]	# Model st. papers [56]	Our findings
∅	*	—	36.5%
<100	2	5	15.5%
100-1k	8	9	4.8%
1k-100k	1	16	7.4%
>100k	—	10	1.1%
∞	11	40	15.6%

Another assumption in backdoor attacks is that practitioners rely on existing models and fine-tune these. We combined the information provided by Cinà et al. [17] about the fine-tuning setting and our participants' replies ($Q21$). Of the 32 backdoor papers, 12 dealt with a fine-tuning setting, e.g., the victim took an external model and fine-tuned this model on internal data. Almost half of our participants (48.1%) stated to use third-party models and then fine-tune them. Only about a quarter denied using any third-party models (24.3%). This setting was studied in 12 (37.5%) of the backdoor papers. These findings highlight the need to study security risks both for pre-trained and end-to-end training, as is currently the case. Furthermore, backdooring or poisoning a model used later on circumvents the need to alter training data.

Discussion. While there are notable exceptions of papers assuming very small poisoning/backdoor percentages of less than 3% in vision [12, 33], object detection [48], and point clouds [74], more such work is needed. Furthermore, there are two limitations to discuss. On the one hand, we currently do not know which quality checks are put upon public data, and how this affects current attacks. In addition, an attack altering 5% of the training data may affect 20.4% of our participant's models, as the true allowed percentage may be lower. However, the attack affects 40.6% of the cases, since *at least* 5% data access is required.

Take away–Training time attacks. We find evidence that assumptions of poisoning and backdoor threat models are met in practice. Yet, while data can often not be accessed directly, poisoning and backdooring may be executed via public data sources. Our participants also reported frequent (about 50%) use of third-party models which are then fine-tuned.

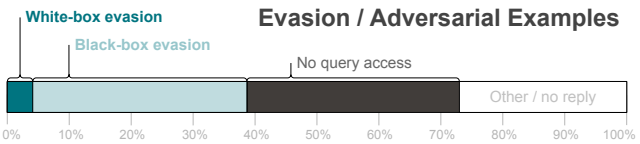

Figure 2: Evasion threat models in percent of our participants' replies. We report 3rd party access: White denotes incomplete data or an irrelevant threat model (e.g., only model accessible). Black represents no access, turquoise white-box and light turquoise black-box evasion threat models.

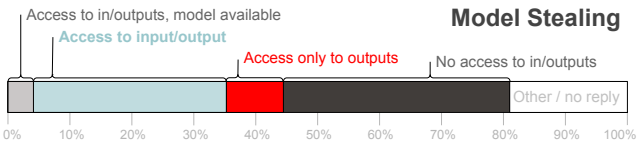

Figure 3: Model stealing threat model in percent of our participants' replies. We describe 3rd party access: White denotes incomplete data or an irrelevant threat model (e.g., only test inputs are accessible). Black represents no access, turquoise denotes the academic threat model, gray that the attack is obsolete as the model is available. Red denotes a rarely studied threat model in current research.

4.2.2 Test-Time Attacks

Evasion, model stealing, and privacy-based attacks target a model at test time (as visualized in Table 2). They are thus similar as they require the submission of test inputs and observing the model's outputs. Before we cover these attacks individually, we examine these requirements in general.

In terms of test data access ($Q33$), almost half (48.1%) of our participants reported that the model could not be queried. On the other hand, 39.5% reported that querying their model was possible. The model itself ($Q24$) was not accessible for three-quarters (75.5%) of our participants, for 7.7% of them, the model was publicly available. Model outputs ($Q32$) were available more readily: outputs were not accessible in roughly a third (37%), and freely available in half (49.1%) of the cases.

To understand how many queries could be submitted at test time ($Q33$), we briefly report statistics. In most (36.5%) cases, no queries were possible. This is followed by less than 100 queries (15.6%) and infinitely many (15.5.%), followed by 1,000 - 100,000 (7.4%). The least frequent are more than 100,000 queries (1.1%) and 100-1,000 queries.

> **Take away–Test time attacks.** Compared to the training data, the threat surface is larger at test-time but queries to accessible models are either very constrained or unconstrained.

To be able to cover each attack's specialties, we now analyze the specific threat models individually.

Evasion. Many evasion attacks assume access to the model and the model's inputs at test-time to alter predictions [9, 19, 28, 49] (see also Table 2). We examined these threat models and visualized our participants' replies in Figure 2. We found that 3rd party access to these two features ($Q24$ and $Q33$) was rare and only reported by 4.1% of our participants. If we dropped the white-box constraint and permitted the attacker to have no access to the model, this percentage increased strongly to 34.6%. As expected, black-box attacks could be carried out more frequently in our sample.

We thus focus on black-box attacks [9, 18, 25, 50] and the number of queries needed for an attack ($Q33$). For the sake of this comparison, we relied on the overview of Mahmood et al. [50] and ignored whether attacks are targeted or untargeted and whether hard or soft labels are required. We report the minimal empirical amount of queries documented by Mahmood et al. [50] in Table 5. Few (2) papers operated in the setting most frequently reported (15.6%) with less than 100 queries allowed. Most papers (8) needed 100-1,000 queries, which is the range least often (4.8%) reported by our participants. One paper required 1,000-100,000 queries, which is slightly more frequent (7.4%). On the other hand, 15.5% of our participants stated to allow infinitely many queries. In this sense, access to AI systems in practice was all-or-nothing, with few test queries or infinitely many. Research, in contrast, focused on the middle amount of queries, possibly as a consequence of decreasing the number of queries needed. An in-depth understanding of the required queries to attack a model is subject to ongoing research [25]. In addition, our work is a call for transferability studies, when neither model nor data are known, as uttered by Sheatsly et al. [64]. Such a setting (attacking only via test data) was most practical according to our participants.

> **Take away–Evasion.** According to our participants, 4.1% of their models were vulnerable against white-box evasion. Often, the model is not available; and either very few or an unconstrained number of queries is granted, whereas research assumes a moderate query number. Finally, in some cases only data can be submitted without model feedback, highlighting the need to deepen our understanding of transferability.

Model stealing. Model stealing attacks target the model via test inputs and outputs [56] ($Q24$, $Q32$, and $Q33$). The goal is to obtain a copy of the target model in terms of functionality or a direct copy of the weights. We examined this threat model and plotted the corresponding percentages in Figure 3. 44.5% of our participants reported that they allowed public access to model outputs. Most model stealing attacks [56, 71] require submitting specific queries, decreasing this percentage to 35.3%. In 4.1% of our sample, the model itself was however also accessible, defeating the purpose of the attack. Although the assumptions of model stealing are met in some

cases, in about 10% of the cases, it would be beneficial to study model stealing attacks that are purely based on observing the outputs of samples that are not under the attacker's control, as somewhat studied by Papernot et al. [57].

An additional factor in model stealing is, as before, the submittable number of queries to the target model ($Q33$). We compared the number of queries reported by Oliynyk et al. [56] to our sample in Table 5. Most of the 40 papers surveyed required between 100 and 100,000 queries, the numbers our participants reported the least frequently. Only five papers relied on less than 100 queries and aligned with a larger (15.6%) percentage within our sample. We further investigate the relationship between the number of queries allowed and model complexity (as approximated by input size, $Q29$). There is no statistically significant correlation. The most frequent combinations of replies were with 10.7% inputs of size 100-1k, 9.3% 10-100, and 6.3% no applicable feature size, each with less than 10 queries. Both an input size of 10-100 with 10-100 queries and not applicable input size with unconstrained inputs were reported in our sample in 4.1%. All other combinations appeared ten times or less in the responses, with 24.4% responses not being analyzed due to missing data.

Discussion. As before, an attack needing $<$100 queries possibly applies to 15.6% of the models within our sample, but also in 28.8% (4.8%+7.4%+1.1%+15.6%, see Table 5) of the cases, as *at least* 100 queries are required.

> **Take away—Model stealing.** Model stealing can be carried out in practice. Yet, in some cases where input and output are accessible, the model is accessible, too. According to our sample, a relevant setting for model stealing attacks is only output visibility, without the possibility of submitting test queries. In addition, most attacks study infrequent numbers of queries, as either more or fewer samples are granted commonly. More work is needed to understand the relationship of amount of queries and model complexity in practice.

4.2.3 Privacy Attacks

We here discuss the two attacks inferring training data properties, first membership inference and then attribute inference.

Membership inference. Membership attacks use test queries and their corresponding output from the target model to infer information about the training data [35, 39] ($Q23, Q32, Q33$; see Table 2). We visualize the practical threat models at the top of Figure 4. About half of our participants allowed 3rd party access to their model outputs. When combined with accessible test data, this decreased to 37.3%. In 4.4% of these cases, the training is then public, too. Independently, most membership attacks [35] assumed one input and output per training point to determine membership. As the number of queries was often less than 100 with only outputs visible,

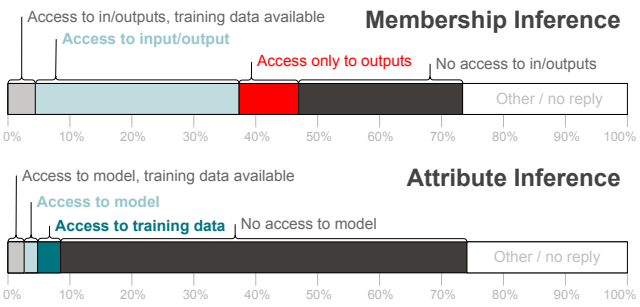

Figure 4: Membership and attribute inference threat models in percent of participants' replies. We describe 3rd party access: White denotes incomplete data or irrelevant threat models, black represents no access, turquoise denotes existing threat models, gray means that the attack is obsolete as the training data is available, too. For membership, red denotes a threat model not studied so far. In the case of attribute inference, turquoise denotes no model access, but the property can directly be inferred from the training data.

it would be beneficial to understand if membership can be inferred for several points at once.

Property inference. Property inference attacks derive from the model's weights properties of the training data [39] ($Q25, Q23$; see Table 2). Analogous to previous observations, 65.7% of the participants give no access and 4.8% grant 3rd party access to their weights. In 2.5% of these cases, however, the training data is then publicly available.

> **Take away—Privacy attacks.** Membership and property inference can be carried out in practice. In some cases where threat models apply, the data was however accessible, too. According to our sample, it would be beneficial to study membership attacks with fewer attacker capabilities: only access to outputs.

4.3 AI Security Beyond Specific Attacks

There is more to learn from our survey respondents than attack-specific threat models. This section presents information that either supports existing work or can be used to support future, realistic AI threat modeling. To this end, we first discuss factors like common dataset sizes, involvement of domain experts, real-time requirements, library usage, and finally which factors influenced given access to AI systems.

4.3.1 Practical Challenges for AI Security Research

To better align AI security research and practice, we discuss relevant information to make future work in AI security more practical. To this end, we review which data types are commonly used in the industry, and then discuss practical challenges on both the attack and defense sides.

Table 6: Summary of our participant's reported data. We denote the sample size in the number of features x and the size of the training set |X, Y|. We also give examples of academic datasets of similar dimensions. 21.8% partial replies are not listed, neither are combinations with a prevalence <5%.

	size of x	|X, Y|	Example dataset
9.6%	10-100	10^5-10^8	
9.6%	10^2-10^3	10^3-10^5	(Fashion) MNIST [45, 75]
7.8%	10-100	10^3-10^5	Iris [3], Wine [3], Spam [3]
7%	10^2-10^3	10^5-10^8	
5.5%	10^3-10^5	10^2-10^3	CIFAR [44], Drebin [2]
5.5%	other	10^3-10^5	
5.2%	other	other	Open AI Gym [11]

Common dataset sizes. We investigated the most frequent dataset properties in terms of feature size ($Q29$) and training set size ($Q30$) in Table 6, as data dimensionality (e.g., number of features) and samples may influence security [73]. Considering both questions in combination, 59 participants (21.8%) did not reply to one of the questions. Overall, the data size was very diverse. Yet, small input sizes (10-100 or 10^2-10^3) were prevalent. Only considering $Q29$ about feature sizes, over fifty percent of our participants reported a small number of features (10-100 (27.2%) or 10^2-10^3 (25.8%)). Also, non-quantifiable data was frequent (16%), as shown in Table 6. These sizes match datasets such as CIFAR [44], CelebA [47], and Open AI Gym [11]. We also found dimensions of datasets that were used heavily before, including MNIST [45], Drebin [2], and smaller datasets like Iris [3], Wine [3], and Spam [3]. Overall, frequent datasets in practice were smaller than current academic datasets: there were no image-net [20] like datasets frequent in our sample. Many practitioners worked thus with smaller data in terms of features (with a potentially large number of samples) which has, to the best of our knowledge, not been studied in depth yet.

Attack and mitigation challenges. We investigate two more properties that potentially affect AI security. First, we asked our participants whether they relied on a domain expert ($Q15$). The presence of an expert may imply that also to attack, specific knowledge is required to constrain for example feature changes. A large fraction (37.8%) of our participants reported relying on domain experts. An additional 4% wanted to work with one but did not find someone yet. Domain knowledge should thus be considered when studying AI security.

Furthermore, we inquired whether our participants required real-time responses to their AI-based systems ($Q17$). Such a requirement affects the overall time that is available to defenses. Only 11% of all participants reported that their applications were *not* time-critical at all, but only about a third (35.8%) required real-time results. This shows the need for mitigations to cope with time constraints practice.

Take away–General threat modelling. Current research datasets match practical settings. Yet, some are in practice smaller in features than current academic counterparts, outlining the need to also study data security for a few features and many samples. Furthermore, our results emphasize the need to study constraints in terms of expert knowledge or time.

4.3.2 Code Libraries as a Security Factor

While libraries are acknowledged as a relevant security factor in other areas than AI security [58], few works study AI security about libraries [16, 31, 37, 66, 67]. We briefly state our results here as to whether our participants used libraries.

Vulnerabilities via AI libraries. A recent strain of attacks relies on manipulated libraries [31, 66] to affect the order of data [66] or the initial weights of a target model [31]. To assess the feasibility of such attacks, we inquired how models are developed in terms of code ($Q19$). While almost all participants (90.5%) developed models using self-written code, they further relied on additional tools. Such tools included open-source code (88.8%) or proprietary solutions (36.5%). While the aforementioned attacks are complex as a tempered library has to be placed on the victim's machine first, it might be worthwhile to investigate whether hashcodes and other security measures are in place to prevent such attacks.

Energy saving libraries. Several recent works increase the run-time of deep learning models under the assumption that these models use energy-saving soft- or hardware [16, 37]. To verify that this is the case also in practice ($Q22$), we inquired about the use of energy-saving libraries, software, or self-written code. More than a third of our participants confirmed using such methods (34.3%), with almost an additional quarter (24%) stating that they sometimes relied on these techniques. Roughly a third (31.4%) reported not to rely on energy saving. With the constraint that to the best of our knowledge, there is currently no understanding of whether attacks transfer between different energy-saving methods, it seems worthwhile to investigate the security of such libraries further.

Take away. Libraries can be security relevant for AI.

4.3.3 Factors Determining Practical Threat Models

Finally, we review potential factors that influence security factors such as access control on training data and model ($Q23$-$Q24$), public fractions of training ($Q28$) and test data ($Q27$), submittable queries ($Q32$), visible outputs ($Q33$) and usage of pre-trained models ($Q21$) and reliance on domain experts ($Q15$). We tested candidate variables like AI knowledge ($Q5$), security knowledge ($Q6$), AI security knowledge ($Q7$), company size ($Q10$), AI maturity ($Q14$), team size ($Q11$) and presence of a domain expert ($Q15$) and computed the

Table 7: Analyzing factors that influence threat model properties. We denote a negative correlation with n and the absence of a statistically significant relationship with –.

	AI Know.	Sec. Know.	AI Sec. Know.	Comp. Size	AI Maturity	Team Size	Domain Expert
Access (X,Y)	–	–	–	–	–	–	–
Access ω	–	–	–	–	–	–	–
# queries	–	–	–	–	–	–	–
Access F (ω, x)	–	–	–	–	–	–	–
(X_t, Y_t) from public	n	–	–	–	n	–	–
(X,Y) from public	n	–	–	–	n	–	–
Pretrained F	–	–	–	–	–	–	–
Domain expert	–	–	–	–	–	–	

Spearman correlation to determine relationships. We set as p value 0.05 with Bonferroni correction for repeated testing, yielding a significance level of $0.05/56 = 0.0009$. Most of these combinations were not statistically significant, with a few notable exceptions, as visible in Table 7.

AI maturity and company size. In our sample, AI maturity affected how much test and training data came from public sources. For both training (-0.24, $p = 4.8e^{-07}$) and test (-0.34, $p = 0.0005$) data, the correlation was negative. A negative correlation indicated that more mature companies tended to collect less data from public sources.

AI security and AI knowledge. AI knowledge affects, analogous to AI maturity, how much training (-0.23, $p = 0.00077$) and test data (-0.23, 0.0008) were sourced from public places. As with AI maturity, this correlation was negative. This indicated that as practitioners were more knowledgeable, less data stemmed in either case from public sources.

Possible influences. A few factors in our survey affected security-relevant features like access to the model, model components, or data. A possible explanation is that these are influenced by factors not considered in this survey, for example, business models, the application, or industry area. In contrast, data collection practices are correlated to AI knowledge and the AI maturity of the company.

> **Take away.** AI security-relevant factors are not correlated to security or AI security knowledge in our sample. Both AI maturity and AI knowledge influence negatively whether data comes from public sources.

5 Limitations

In this section, we discuss the limitations of our study. We first describe sample limitations, then proceed to discuss limitations within our questionnaire, and conclude the section by discussing methodological limitations.

Sample limitations. Our sample is biased towards the global north, especially Europe, and is limited to English-speaking practitioners. Albeit we managed to recruit over 250 participants, we could not find reliable and consistent scientific references to estimate the global target population of industrial practitioners working with AI. However, for a population larger than 50,000, and a confidence interval of 95%, our sample's margin of error lies around 6%. Reducing this margin significantly to a few percent, for example, 2%, would require several thousand participants. Furthermore, in terms of demographics, our sample matches the overall population [42] rather well (Sect. 3.4). Given that our goal is to identify conceptual mismatches of threat models in the wild compared to research, we find this margin of error acceptable.

Questionnaire limitations. Despite our best efforts and many pretests, some questions could not be used for our analysis. This included information on the detectability of attacks ($Q40$) and questions where we inquired information about the secrecy of the input encoding ($Q18$), the expected performance ($Q34$), and the ease to assess the quality of training data ($Q35$). These questions would have helped to determine the difficulty an attacker faces when targeting a model. We relied on responses on a scale from 1 to 100. Still, the distribution of replies matched a normal distribution with a mean of 50 and quartiles around 25 and 75, indicating that the questions did not contain enough information for analysis. We leave a detailed study of these aspects for future work. Orthogonally, we had planned to ask for security-relevant output transformations like not providing confidence scores. However, our pre-tests outlined that this was too specific for a sample covering RL or DM with non-discrete outputs. We thus left a study of these aspects for future work.

Methodological limitations. We did not review the entire body of AI security work. Given that there are several thousand research articles about AI security[3], this endeavor is beyond a single paper. We instead rely on surveys [16, 35, 39, 50, 56] representing the state of the art for different attacks. We chose these surveys explicitly as they reviewed properties related to the threat models of the analyzed attacks. Some of these surveys focus on specific areas like computer vision [17]. The scope of our comparison is thus limited and may be biased. Yet, we reason that this overview is sufficient to identify conceptual gaps. In addition, some attacks depended on factors like memorization or overfitting (membership inference) [78]. As these are complex phenomena, we opted against analyzing them. While this limits our insights on these attacks, we leave this aspect for future work. Orthogonally, it is important to recognize that our study relies on self-reported properties and sheds light on what attacks are

[3] https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

possible through channels known to practitioners. Real-world practical attack threat levels may be higher.

Independently, the practical threat models we discuss represent a momentary picture of how AI is applied in practice. Usage may change over time, resulting in evolving threat models, which should be monitored over time. Finally, AI usage is strongly dependent on the context of a specific application, which we do not cover but leave for future work.

6 Implications and Future Work

Having discussed the limitations, we are ready to discuss the implications and implied future work of our study. As the most important implication of our work is directing future research in AI security, we first discuss these research directions. Afterwards, we discuss additional implications, concerning AI regulation and AI security in practice. Where applicable, we also delve into future work for these latter implications.

6.1 Future Work in AI Security

We found several gaps between the researched threat models and practical AI usage (Sect. 4). Consequently, most of our implications translate to direct recommendations of previously overlooked aspects. In this section, we give the big picture by combining our findings for each attack, listing open questions alongside. An overview of these results can be found in Table 8. At the end of the section, we summarize insights that go beyond individual attacks.

Poisoning and backdoors. Poisoning and backdoor threat models apply in practice (Sect. 4.2.1). Further studies should focus on ending the arms-race and deepening our knowledge of defense trade-offs [17]. At the same time, current percentages of frequently altered training data are not well aligned with the percentages reported by our practitioners (Sect. 4.2.1). Although some practitioners currently report high training amounts from public sources, this is deemed to decrease as attacks or data quality problems occur. Finally, given that practitioners rely on fine-tuned pre-trained AI models (Sect. 4.2.1), corresponding risks need to be assessed [36].

Evasion. We found evidence of the applicability of (black-box) evasion threat models (Sect. 4.2.2), and recommend further study to end the arms-race [18, 70]. Still, more emphasis should be put on studying attacks that succeed without knowledge of the exact data and model outputs (Sect. 4.2.2). This is aligned with previous observations that "attackers don't compute gradients" [1], as for gradients both input and output pairs are required. A similar perspective on this requirement is that more work is required on transferability. More precisely, and as stated by Sheatsly et al. [64], more work should study transferability across different datasets, not only across models. If queries are allowed, the number of queries should be minimized, ideally to less than 100, to reflect frequent settings within our sample (Sect. 4.2.2). Frequently, there were also no limitations on the number of queries. Yet, adding such constraints is straightforward, and security assessments should not rely on changeable configurations.

Model stealing. We found evidence of the applicability of model-stealing threat models (Sect. 4.2.2). Future work should address the corresponding arms-race [56]. We further found a mismatch of used queries in model stealing and a mismatch for the attacker's capabilities overall (Sect. 4.2.2). Consequently, we recommend reducing used queries, and not relying on currently reported high amounts of queries, similar to evasion. In some cases, only outputs are observable in our sample. It may thus be beneficial to understand the limitations of retrieving information only by observing outputs [57]. In addition, more work should study how query number and model complexity relate in practice. Such results would also hold implications for other inference attacks based on test queries like inversion attacks [39] or model extraction [39].

Membership and attribute inference. We found evidence of the applicability of membership and property inference threat models in practice. We should thus address the ongoing arms race to defend such threats [39]. Within our sample, the threat model for attribute inference was rare (Sect. 4.2.3). In membership, we recommend investigating attacks being staged without control over the submitted test queries (Sect. 4.2.3). For both attacks, more understanding of minimal knowledge attacks would be beneficial, for example, to infer membership for several points from only one query.

Security relevance of libraries. We found evidence that energy-saving libraries are frequently ($>33\%$) used in practice (Sect. 4.3.2). It would thus be beneficial to study sloth attacks, in particular with a focus on different energy-saving approaches, and whether attacks transfer across them or not. In addition, it is important to comply upfront with the corresponding constraints of practical training (Sect. 4.2.1) and test-time (Sect. 4.2.2) threat models.

Attack cost and stealthiness. There are possible costs attached to querying and changing data. A general focus on attacks with very few required resources is thus beneficial to understanding real-world vulnerabilities. Another aspect is the stealthiness of the altered data, which is already object to debate [27]. This stealthiness may also be related to domain knowledge (Sect. 4.3.1), where more work is required to understand the nature of these constraints and how frequently they occur in practice. This is also loosely related to the question of whether and to what degree the attacker needs to know the exact data distribution of the victim. Finally, stealthiness needs to be studied in relation to human perception [23], but also in the context of the limitations of automated detection.

Table 8: **Main results.** All attacks exist in practice (∃?) and require mitigations. We also list the threat models that practice-oriented research should focus on and denote the prevalence in our sample, where a higher prevalence may indicate a practically more relevant attack. We then summarize additional directions for research.

Attack	∃?	Relevant Practical setting to be researched	Prevalence in sample	Possible further research
Poisoning [17]	✓	<10% training data alterable	rare	defense trade-offs
Backdoor [17]	✓	<10% training data alterable	rare	model re-use
Evasion, white-box [68]	✓		rare	(model) transferability
Evasion, black-box [50]	✓	<100 queries possible	> 33%	(model) transferability
Model Stealing [56]	✓	<100 queries	> 33%	attacks without query access
Mem. Inf. [39]	✓	<100 queries	> 33%	attacks without query access
Attribute Inf. [39]	✓		rare	attacks without model access

6.2 Practical Implications

Our research has implications beyond AI security research, which we discuss now. The most important implication is that our anonymous participants' AI systems may be vulnerable, highlighting the need for deployable, practical mitigations.

Regulatory and societal implications. Assessing the true vulnerability of AI systems in practice is required by legislative approaches and regulatory frameworks such as the EU AI Act. With poisoning and evasion, Article 15 of the latter even explicitly names some threat models whose underlying assumptions we could confirm (Sect. 4). Technical solutions and organizational measures to address such vulnerabilities are relevant. However, our study shows that the legal text's addition 'where appropriate'[4] is crucial. Not every threat model applies for each AI system as access schemes vary (Table 3) and may be related to use-case, industry area, and other factors, which are left for future work. We also took a first step towards understanding what influences security-relevant features of models (Sect. 4.3.3). More work is needed here as well. Yet, our results show that security is not a primary influence, implying that to make AI systems more secure, regulations are needed to prevent possible future incidents. Beyond regulation, assessing vulnerabilities helps to manage the risk of potential security incidents. Using our threat models (Sect. 4), the risk assessment of AI products in practice can now be completed as previously unknown settings can be studied. In this sense, our work has the potential to reduce what formerly were blind spots in AI systems.

AI security in practice. All 6 attacks studied within the framework of AI security are theoretically possible in practice. The low reported percentage of vulnerable settings reported within our sample is however rather small, potentially contributing to an explanation of few found AI security incidents [30]. Furthermore, although common academic dataset specifications do occur in practice, many participants reported small feature sizes and large numbers of samples (Sect. 4.3.1). Understanding the effect of a small feature space with a potentially large amount of training is thus required. Analogously, we need to understand the limitations of not knowing data in practice. In tasks such as malware detection, feature encodings are secret, limiting the attacker [9]. More work is needed to understand these limitations and how frequent they are in practice. Orthogonally, we recommend more work studying what influences the exact configuration of threat models in organizational contexts (Sect. 4.3.3). A deep understanding of which threat models are used in which cases could help to anticipate and mitigate vulnerabilities, but also understand which properties enable vulnerabilities in the first place.

[4] https://data.consilium.europa.eu/doc/document/ST-5662-2024-INIT/en/pdf

7 Related Work

While several contributions criticize existing AI security threat models [1, 16, 24, 27], to the best of our knowledge, no other work provides an overall picture of this research gap.

The closest related work to this study is a questionnaire-based quantitative study by Grosse et al. [30]. They study AI practitioners' AI security perception, asking about practitioners' general concerns and individual attacks. Grosse et al. also used statistical tests to investigate influences on attack concerns. In our work, in contrast, we study whether academic threat models and AI usage in practice are aligned, and thus if AI security concerns are justifiable.

Several works collect loosely similar information as us, including Kaggle's annual report about ML and data science [42], which provides information on, for example, the algorithms used in practice. Furthermore, Nahar et al. [54] investigated the origin of the used data in a small qualitative sample and the data engineer's effect on data requirements. Dilhara et al. [21] studied the usage of libraries in ML-based code within public repositories, not industry applications. Renieris et al. [60] examined the practical usage of third-party

tools and found that almost three-quarters use such tools. The same authors [60] show that such tools may cause AI failures. Finally, Mink et al. [53] investigate why the number of deployed AI security mitigations in practice is rather low.

Previous works reported low [10] to medium [30] AI security concern by industrial practitioners—our work indicates that this impression may stem from an indeed small attack surface due to little granted access to AI models in practice.

8 Conclusion

We took a significant step towards more practical AI security research. We surveyed common threat model properties in practice and matched these to 6 threat models from AI security research. Our findings have implications for current legislative attempts like the EU AI Act that require security and vulnerability assessments of AI systems. We also set previously low numbers of AI security incidents into context, although our sample provides only initial insights on vulnerabilities through 3rd party accessible channels. Real-world vulnerabilities may be higher. Our work also paves the way toward a deep understanding of the security of AI-based products in practice. Most importantly, while academia, despite criticism, has elaborated valid threat models, we also identify significant gaps. Current threat models are too generous about, for example, training data access or test time queries. More practical threat models should be researched. At the same time, attacks requiring few resources can potentially be applied to a large fraction of models. A black-box evasion attack with less than 100 queries, for example, could target 28.8-44.4% of the models in our sample. Even if just a small portion of companies match the exact constraints assumed in academic papers, these systems have to be defended.

Acknowledgments

We thank our participants, Hyrum Anderson, Marielle Dado, Daryan Dehghanpisheh, Nikki Hogg, Ritesh Sharma, Aryan Trip, Karn Wong, Alla Zhdan, and the MLOps community for their support. We also thank the anonymous reviewers and our shepherd for their valuable feedback.

References

[1] Giovanni Apruzzese, Hyrum Anderson, Savino Dambra, David Freeman, Fabio Pierazzi, and Kevin Roundy. Position:"real attackers don't compute gradients": Bridging the gap between adversarial ml research and practice. In *IEEE SaTML*. IEEE, 2022.

[2] Daniel Arp, Michael Spreitzenbarth, Malte Hubner, Hugo Gascon, and Konrad Rieck. Drebin: Effective and explainable detection of android malware in your pocket. In *NDSS*, pages 23–26, 2014.

[3] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[4] Giuseppe Ateniese, Luigi V Mancini, Angelo Spognardi, Antonio Villani, Domenico Vitali, and Giovanni Felici. Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers. *Int. Journal of Sec. and Networks*, pages 137–150, 2015.

[5] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D Joseph, and J Doug Tygar. Can machine learning be secure? In *CCS*, pages 16–25, 2006.

[6] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *Int. Conf. on Machine Learning and Data Mining in Pattern Recognition*, pages 262–275. Springer, 2017.

[7] Lukas Bieringer, Kathrin Grosse, Michael Backes, and Katharina Krombholz. Mental models of adversarial machine learning. In *SOUPS*, pages 97–116, 2022.

[8] Battista Biggio, Blaine Nelson, and Pavel Laskov. Support vector machines under adversarial label noise. In *ACML*, 2011.

[9] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

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

[11] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv:1606.01540*, 2016.

[12] Nicholas Carlini. Poisoning the unlabeled dataset of semi-supervised learning. In *USENIX Security*, 2021.

[13] Kangjie Chen, Shangwei Guo, Tianwei Zhang, Xiaofei Xie, and Yang Liu. Stealing deep reinforcement learning models for fun and profit. In *Asia CCS*, 2021.

[14] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv:1712.05526*, 2017.

[15] Christopher A Choquette-Choo, Florian Tramer, Nicholas Carlini, and Nicolas Papernot. Label-only membership inference attacks. In *ICML*, pages 1964–1974. PMLR, 2021.

[16] Antonio Emanuele Cinà, Ambra Demontis, Battista Biggio, Fabio Roli, and Marcello Pelillo. Energy-latency attacks via sponge poisoning. *arXiv:2203.08147*, 2022.

[17] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Comput. Surv.*, pages 1–39, 2023.

[18] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. In *NeurIPS Datasets and Benchmarks Track*, 2021.

[19] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. Adversarial classification. In *KDD*, pages 99–108, 2004.

[20] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE CVPR*, pages 248–255, 2009.

[21] Malinda Dilhara, Ameya Ketkar, and Danny Dig. Understanding software-2.0: A study of machine learning library usage and evolution. *ACM Trans. on Software Eng. and Methodology (TOSEM)*, pages 1–42, 2021.

[22] Khoa Doan, Yingjie Lao, Weijie Zhao, and Ping Li. Lira: Learnable, imperceptible and robust backdoor attacks. In *IEEE ICCV*, pages 11966–11976, 2021.

[23] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *NeurIPS*, 2018.

[24] Ivan Evtimov, Weidong Cui, Ece Kamar, Emre Kiciman, Tadayoshi Kohno, and Jerry Li. Security and machine learning in the real world. *arXiv:2007.07205*, 2020.

[25] Washington Garcia, Pin-Yu Chen, Hamilton Scott Clouse, Somesh Jha, and Kevin RB Butler. Less is more: Dimension reduction finds on-manifold adversarial examples in hard-label attacks. In *IEEE SaTML*, pages 254–270, 2023.

[26] Jonas Geiping, Liam H. Fowl, W. Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *ICLR 2021*, 2021.

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

[28] Abhiram Gnanasambandam, Alex M Sherman, and Stanley H Chan. Optical adversarial attack. In *ICCV*, pages 92–101, 2021.

[29] Maziar Gomrokchi, Susan Amin, Hossein Aboutalebi, Alexander Wong, and Doina Precup. Membership inference attacks against temporally correlated data in deep reinforcement learning. *IEEE Access*, 2023.

[30] Kathrin Grosse, Lukas Bieringer, Tarek R Besold, Battista Biggio, and Katharina Krombholz. Machine learning security in industry: A quantitative survey. *IEEE Transactions on Inf. Forensics and Sec.*, pages 1749–1762, 2023.

[31] Kathrin Grosse, Thomas A Trost, Marius Mosbach, Michael Backes, and Dietrich Klakow. On the security relevance of initial weights in deep neural networks. In *ICANN*, pages 3–14, 2020.

[32] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the ml model supply chain. *arXiv:1708.06733*, 2017.

[33] Xingshuo Han, Guowen Xu, Yuan Zhou, Xuehuan Yang, Jiwei Li, and Tianwei Zhang. Physical backdoor attacks to lane detection systems in autonomous driving. In *ACM Int. Conf. on Multimedia*, pages 2957–2968, 2022.

[34] Paul A Harris, Robert Taylor, Brenda L Minor, Veida Elliott, Michelle Fernandez, Lindsay O'Neal, Laura McLeod, Giovanni Delacqua, Francesco Delacqua, Jacqueline Kirby, et al. The redcap consortium: building an int. community of software platform partners. *Journal of biomedical informatics*, 95:103208, 2019.

[35] Xinlei He, Zheng Li, Weilin Xu, Cory Cornelius, and Yang Zhang. Membership-doctor: Comprehensive assessment of membership inference against machine learning models. *arXiv:2208.10445*, 2022.

[36] Sanghyun Hong, Nicholas Carlini, and Alexey Kurakin. Handcrafted backdoors in deep neural networks. *NeurIPS*, pages 8068–8080, 2022.

[37] Sanghyun Hong, Yiğitcan Kaya, Ionuţ-Vlad Modoranu, and Tudor Dumitraş. A panda? no, it's a sloth: Slowdown attacks on adaptive multi-exit neural network inference. *arXiv preprint arXiv:2010.02432*, 2020.

[38] Information technology—Artificial intelligence—Guidance on risk management. Standard, Int. Organization for Standardization, Geneva, CH, March 2023.

[39] Marija Jegorova, Chaitanya Kaul, Charlie Mayor, Alison Q O'Neil, Alexander Weir, Roderick Murray-Smith, and Sotirios A Tsaftaris. Survey: Leakage and privacy at inference time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[40] Yujie Ji, Xinyang Zhang, and Ting Wang. Backdoor attacks against learning systems. In *IEEE CNS*, pages 1–9, 2017.

[41] Matthew Joslin, Neng Li, Shuang Hao, Minhui Xue, and Haojin Zhu. Measuring and analyzing search engine poisoning of linguistic collisions. In *IEEE S&P*, 2019.

[42] Kaggle. State of machine learning and data science. https://www.kaggle.com/kaggle-survey-2021, 2021.

[43] Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li. Trojdrl: evaluation of backdoor attacks on deep reinforcement learning. In *DA Conf.*, pages 1–6. IEEE, 2020.

[44] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[45] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[46] Yen-Chen Lin, Zhang-Wei Hong, Yuan-Hong Liao, Meng-Li Shih, Ming-Yu Liu, and Min Sun. Tactics of adversarial attack on deep reinforcement learning agents. In *IJCAI*, pages 3756–3762, 2017.

[47] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[48] Hua Ma, Yinshan Li, Yansong Gao, Alsharif Abuadbba, Zhi Zhang, Anmin Fu, Hyoungshick Kim, Said F Al-Sarawi, Nepal Surya, and Derek Abbott. Dangerous cloaking: Natural trigger based backdoor attacks on object detectors in the physical world. *arXiv:2201.08619*, 2022.

[49] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

[50] Kaleel Mahmood, Rigel Mahmood, Ethan Rathbun, and Marten van Dijk. Back in black: A comparative evaluation of recent state-of-the-art black-box attacks. *IEEE Access*, 10:998–1019, 2021.

[51] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Adversarially robust policy learning: Active construction of physically-plausible perturbations. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 3932–3939. IEEE, 2017.

[52] Shike Mei and Xiaojin Zhu. The security of latent dirichlet allocation. In *Artificial Intelligence and Statistics*, pages 681–689. PMLR, 2015.

[53] Jaron Mink, Harjot Kaur, Juliane Schmüser, Sascha Fahl, and Yasemin Acar. "security is not my field, i'm a stats guy": A qualitative root cause analysis of barriers to adversarial machine learning defenses in industry. In *USENIX Security*, 2023.

[54] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. *Organization*, 1(2):3, 2022.

[55] Seong Joon Oh, Bernt Schiele, and Mario Fritz. Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 121–144, 2019.

[56] Daryna Oliynyk, Rudolf Mayer, and Andreas Rauber. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys*, 2023.

[57] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Asia CCS*, pages 506–519, 2017.

[58] Gede Artha Azriadi Prana, Abhishek Sharma, Lwin Khin Shar, Darius Foo, Andrew E Santosa, Asankhaya Sharma, and David Lo. Out of sight, out of mind? how vulnerable dependencies affect open-source projects. *Emp. Software Eng.*, pages 1–34, 2021.

[59] Aravind Rajeswaran, Sarvjeet Ghotra, Balaraman Ravindran, and Sergey Levine. Epopt: Learning robust neural network policies using model ensembles. In *ICLR*, 2017.

[60] Elizabeth M Renieris, David Kiron, and Steven Mills. Building robust rai programs as third-party ai tools proliferate. *MIT Sloan Management Review*, 2023.

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

[62] Alex Serban and Joost Visser. Adapting software architectures to machine learning challenges. In *IEEE Int. Conf. on Software Analysis, Evolution and Reengineering (SANER)*, pages 152–163, 2022.

[63] Ali Shafahi, W. Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *NeurIPS*, 2018.

[64] Ryan Sheatsley, Blaine Hoak, Eric Pauley, and Patrick McDaniel. The space of adversarial strategies. In *USENIX Security*, pages 3745–3761, 2023.

[65] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against ml models. In *IEEE S&P*, pages 3–18, 2017.

[66] Ilia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A Erdogdu, and Ross J Anderson. Manipulating sgd with data ordering attacks. *NeurIPS*, 34:18021–18032, 2021.

[67] Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, and Ross Anderson. Sponge examples: Energy-latency attacks on neural networks. *arXiv:2006.03463*, 2020.

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

[69] Te Juin Lester Tan and Reza Shokri. Bypassing backdoor detection algorithms in deep learning. In *EuroS&P*, 2020.

[70] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *NeurIPS*, pages 1633–1645, 2020.

[71] Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *USENIX*, 2016.

[72] Kelce S Wilson and Müge Ayse Kiy. Some fundamental cybersecurity concepts. *IEEE access*, 2:116–124, 2014.

[73] Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. *NeurIPS*, 2018.

[74] Zhen Xiang, David J Miller, Siheng Chen, Xi Li, and George Kesidis. A backdoor attack against 3d point cloud classifiers. In *ICCV*, pages 7597–7607, 2021.

[75] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv:1708.07747*, 2017.

[76] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning? In *ICML*, pages 1689–1698. PMLR, 2015.

[77] Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *ACM CCS*, pages 225–240, 2019.

[78] Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Sec. Foundations Symposium (CSF)*, pages 268–282, 2018.

Notes

We depict our study's full questionnaire.

I - Demographics.

$Q1$: How old are you? [18-24, 25-34, 35-44, 45-54, 55-64, 64+]

$Q2$: What gender do you identify with? [Female, male, other, I do not want to disclose]

$Q3$: In which country are you located? Please use the country calling code of your country to choose a group, where X can be replaced by any digit.
$[+1, +299]$ (North America)
$[+2X]$ (Africa)
$[+30 - +35X, +39]$ (Southwest Europe)
$[+36X - +38X]$ (East Europe)
$[+4X]$ (Central / Northern Europe)
$[+52, +53, +50X]$ (Central America)
$[+51, > +54]$ (South America)
$[+6]$ (Oceania, Australia, New Zealand)
$[+7, +976]$ (Russia, north Asia)
$[+8]$ (East Asia, Japan)
$[+90, +96X, +970 - +974]$ (Near East and Türkiye)
$[+91 - +95, +99X, +975, +076, +977]$ (Southeast Asia)

$Q4$: What is your level of education? Please specify the highest. [Highschool, Bachelor, Master / Diploma, Training / Apprenticeship, PhD, Other]

The four following questions (5-8) have all the same replies, namely: [None, Education only, $<$ 1 year, 1-2 years, 2-5 years, 5-10 years, $>$ 10 years]

How many years have you worked in/with...

$Q5$: AI/ML? $Q6$: Security? $Q7$: AI/ML Security?

$Q8$: AI/ML policies or AI/ML risk management?

$Q9$: In which country is your organization headquartered? [replies as above in $Q3$]

$Q10$: What is the number of employees at your organization? [0-49, 50-249, 250-999, $>$1000]

$Q11$: What is the size of the team you work in? [$<$3, 3-5, 6-9, 10-15, $>$15]

$Q12$: In which industry area does your company operate? [Mining, utilities, construction, manufacturing, wholesale trade, resale trade, transport. & warehousing, information, finance & insurance, real estate & rental & leasing, professional & scientific & tech. services, management of companies & enterprises, administration, education, health care & social assistance, arts & entertainment & recreation, accommodation & food services, public administration, other services]

$Q13$: To encompass specific industries, please tick which of the following areas you work in (feel free to tick several)? [academic research, automotive or suppliers of automotive, cyber security, healthcare, none of these]

Part II.A - Your AI based Projects (33% of survey done).

$Q14$: What is the status of the ML projects you work on? [Indirect usage (e.g. certification, auditing); Evaluating use cases; Starting to develop models; Getting developed models

into production, Models in production, for 1-2 years; M. in production, for 2-4 years; M. in production, for >5 years]

In case you work on several projects, from here on, please stick to one (for example your favorite) AI project.

$Q15$: Are you collaborating with domain experts on our data? As an example, consider a healthcare application where a doctor or domain expert needs to be involved. [yes; no, not required; no, none available]

$Q16$: How clear is the specification (in terms of intended use, infrastructure, deployment, etc) of a requested model to you? [linear scale from 0 (very unclear) to 100 (very clear)]

$Q17$: How time-critical is obtaining a result for a query of your AI-based application? [not time-critical at all; time-critical, but not real-time; real-time required]

$Q18$: Are your inputs/features secret because they are customized or hand-engineered? For example, your company-developed representation of a program would be secret, whereas an RGB image encoding would be well-known. [linear scale from 0 (very secret) to 100 (well known)]

Part II.B - Development of AI (45% of survey done).

$Q19$: Which of the following resources do you use to develop your models in terms of code? (several replies are fine) [self written code; open source code; proprietary solutions]

$Q20$: Where do you train your models (several replies are fine)? [on-premise servers; cloud-provided servers; mix of both/hybrid cloud]

$Q21$: Do you use pre-trained third-party models, in other words, models not trained on your own data? (several replies are fine) [yes; yes, but we fine-tune them; yes, but for development only; not in deployment; no]

$Q22$: Do you use any libraries, software, or self-written code to decrease runtime or execution of your models at deployment? Examples would be quantization, ASIC sparsity-based models, etc. [yes; sometimes; no]

Part II.C - Data and Model within your AI-based Project (55% of survey done).

The four following questions (23-26), have all the same replies, namely: [Accessible to 3rd party; under access control; not accessible at all]

Please specify the accessibility of different parts of your ML pipeline to third parties.

$Q23$ Training data. $Q24$ Model (parameters).

$Q25$ Test data. $Q26$ Model outputs.

For the next questions, educated guesses for the responses are sufficient.

$Q27$: Which fraction of your test data comes from public sources (e.g., Internet)? [None; <1%; 1% -5%; 5% -10%; 10% -15%; 25% -50%; 50% -75%; >75%; I don't know]

$Q28$: Which fraction of your training data comes from public sources (e.g., Internet)? [None; <1%; 1%-5%; 5%-10%; 10%-15%; 25%-50%; 50%-75%; >75%; I don't know]

$Q29$: What is the size of your input (e.g., number of features)? [< 10; 10 - 100; 100-1K; o 1k-100K; 100k-110M; >100M; does not apply]

$Q30$: How many samples do you train on? [<10, 10-100; 100-1K; 1k-100K, 100k-100M, >100M; does not apply]

$Q31$: How many samples do you evaluate your model on? [<10, 10-100; 100-1K; 1k-100K, 100k-100M, >100M; stream of data]

$Q32$: How many model outputs can a third party observe? [<10, 10-100; 100-1K; 1k-100K, 100k-100M, >100M; unconstrained]

$Q33$: How many model outputs can a third party query from your model during the entire time the model is available? [<10, 10-100; 100-1K; 1k-100K, 100k-100M, >100M; unconstrained]

$Q34$ Are you able to estimate the performance (expected accuracy) before training? As an example, a very clear case is a known classification with documented accuracy >90%. [linear scale from 0 (not at all) to 100 (absolutely)]

$Q35$ How easy is it for you to assess the quality of your training data? As an example, is the data easy to inspect visually, or can you test whether it corresponds to your task? [linear scale from 0 (not at all) to 100 (absolutely)]

$Q36$ How much can you influence or enforce requirements (such as sampled from a certain source, inspected by a real worker, etc) on the used training data? [linear scale from 0 (not at all) to 100 (absolutely)]

III - AI security (88% of survey done).

$Q37$: How relevant is the security of your AI-based product to you? [1 (very low) to 100 (very high)]

$Q38$: How relevant is the user's privacy of your AI-based product to you? [1 (very low) to 100 (very high)]

$Q39$: How high do you estimate the risk of becoming a victim of an attack related to your AI-based workflows, products, or systems within the next 12 months? [1 (very low) to 100 (very high)]

$Q40$: How likely do you estimate the probability of noticing an attack on your AI-based workflows? [1 (very low) to 100 (very high)]

$Q41$-$Q43$ as in Grosse et al. [30].

$Q41$: Did you already experience a circumvention of your AI-based workflows, products or systems? [yes/no]

IF YES: $Q42$: How many circumventions of your AI-based workflows, products or systems have you experienced? [1,2,3,4,>4]

$Q43$: Please describe the most severe circumvention of your AI-based workflows, products or systems. [text field]