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
kathrin.grosse@epfl.ch

Abstract

Recent works identified a gap between research and practice in artificial intelligence security: threats studied in academia do not always reflect practical use and real-world security risks. For example, while models are often studied in isolation, they form part of larger ML pipelines in practice. Latest works also brought forward that adversarial manipulations introduced by academic attacks are impractical. In this short paper, we take a first step toward describing this disparity. To this end, we revisit the threat models of three 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 also 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 (Barreno et al. 2006; Biggio and Roli 2018; Chen et al. 2017; Cinà et al. 2023; Dalvi et al. 2004; Gu, Dolan-Gavitt, and Garg 2017; Ji, Zhang, and Wang 2017; Oh, Schiele, and Fritz 2019; Papernot, McDaniel, and Goodfellow 2016; Szegedy et al. 2014; Tramèr et al. 2016). Although the attacks studied in these works have been established, increasing criticism targets the threat models used. For example, most academic papers focus on standalone models (Chen et al. 2017; Dalvi et al. 2004; Gu, Dolan-Gavitt, and Garg 2017; Ji, Zhang, and Wang 2017; Oh, Schiele, and Fritz 2019; Papernot, McDaniel, and Goodfellow 2016; Szegedy et al. 2014; Tramèr et al. 2016), while models in practice are generally embedded into pipelines or larger systems (Evtimov et al. 2020; Bieringer et al. 2022). In addition, it has been pointed out that attacks in practice do currently not match the degree of complexity inherent to academic publications (Apruzzese et al. 2022; Grosse et al. 2023b). Also, the measurement of the attacker's manipulations was deemed impractical (Gilmer et al. 2018; Apruzzese et al. 2022), and the overall amount of data available to the attacker (Cinà et al. 2023; Grosse et al. 2023b).

For example, backdoor attacks (Ji, Zhang, and Wang 2017; Cinà et al. 2023) require manipulating the training data. Grosse et al. (2023b) reported cases of training data attacks 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 (Cinà et al. 2023), limiting their usefulness. Analogously, evasion attacks were reported in the wild (Grosse et al. 2023b). Evasion requires the submission of at least one perturbed test sample (Dalvi et al. 2004; Szegedy et al. 2014). Yet the number of AI systems in practice where this is possible is again unknown.

These works illustrate individual mismatches and demonstrate that some aspects of threat models are unaligned between research and practice. The underlying problem, an absence of knowledge on how AI is used in practice, is however still unaddressed. In other words, the underlying problem remains: whether researched threat models are actually *representative* of AI usage in practice. In this work, we take a first step towards measuring this mismatch of AI security research and practice on both the attack level and beyond.

Contributions. In this short version of our long paper (Grosse et al. 2023a), we focus on three commonly studied attacks: backdoors (Cinà et al. 2023), evasion or adversarial examples (Dalvi et al. 2004; Szegedy et al. 2014), and model stealing (Tramèr et al. 2016) and describe them in Sect. 2. To match these threat models with real-world AI usage, we designed a questionnaire and recruited a sample of **271** industrial participants. Our results (Sect. 3) show that all three analyzed attacks from academia are relevant in practice. However, in practice, access to models, queries, and data follows an all-or-nothing principle, where existing academic threat models cover only a fraction of the cases. This research gap does not only consist of too generous assumptions about access to training data and test queries. Instead, in some cases where for example model stealing is possible, the model is public as well, superseding the attack in its current threat modelling.

After revisiting the limitations of our approach (Sect. 4), we discuss the implications of our study (Sect. 5). These go

Table 1: A summary of threat modeling for AI security. Below, we list the attacker's knowledge and capabilities. 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 ($F(\omega, x)$). * indicates that either both marked or one of two properties is required. For all properties, we denote required (●), sometimes required (◐), and not required (○).

	Knowledge			Capabilities			
	X,Y	X_t,Y_t	ω	x	y	x_t	$F(\omega, x)$
Backdoor (Cinà et al. 2023)	●	●	◐	●	◐	●	○
Evasion (Mahmood et al. 2021)	○	●	◐ *	○	○	●	◐ *
Model Stealing (Oliynyk et al. 2023)	○	◐	○	○	○	◐	●

beyond the above-discussed shortcomings of existing, academic threat models. Implications relate also 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 the security of AI-based products in practice. We then review related work (Sect. 6) and conclude our contributions (Sect. 7).

Remark. *This work is not a finger-pointing exercise. So far, AI security research has relied on best practices of security threat modeling, and we confirm that all 3 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 Methodology

In this section, we provide our terminology, the questionnaire and recruiting strategy, and the resulting sample. More details can be found in Grosse et al. (2023a).

2.1 Terminology and Definitions

AI security studies the effect an attacker can have on an existing system or program. We, thus, first define the different attacks and threat model components we consider. We summarize all attacks' threat models in Table 1. In this table, we distinguish the knowledge (e.g., what the attacker knows) and their capability (e.g., what they can affect or alter).

Backdoor attacks alter the training data, (samples and/or labels), with the goal of implementing a backdoor pattern that can be used at test time (Barreno et al. 2006; Cinà et al. 2023). Constraints are often formulated in terms of percentages of training data that is altered.

Evasion attacks alter, at test time, the input data slightly to change the output of the model (Dalvi et al. 2004; Chakraborty et al. 2021). Constraints are often formulated in terms of queries that can be submitted to the model when no access to the model is given (black-box attacks).

Model stealing copies the model without the owner's consent by submitting tailored inputs to a model and observing the corresponding outputs (Tramèr et al. 2016; Oliynyk, Mayer, and Rauber 2023). Analogously to evasion, constraints are often formulated in terms of query numbers that can be submitted to the model.

2.2 Questionnaire and Recruiting

To understand AI security risks in practice, we opt for a quantitative questionnaire. Although we focus on AI model access patterns, we nonetheless query some demographic information to compare to other samples (age, gender, geographic location, company size, industry area, team size, and whether and how long the AI system was in production). Afterward, we inquired about the accessibility of the participants' training and test data, the model, and model outputs as well as how many queries could be submitted to a model. We opted for an anonymous, unpaid questionnaire containing only multiple-choice questions. All fields could be left blank to allow for confidentiality. The full questionnaire is available in Grosse et al. (2023a).

After obtaining approval from our institution's ethical review board, we implemented the questionnaire using RedCap (Harris et al. 2019) and conducted pretests. We then recruited via Slack, personal email, and LinkedIn from April 2023 until July 2023, looking for AI engineers or personnel working on a technical level with AI. We expected that these were the most likely to know about model and data access.

2.3 Sample Description

Of our 271 participants, three-quarters (76%) were male, 18.1% female, and the remainder did not reply or disclose their gender. This ratio is comparable to similar studies (Grosse et al. 2023b) (71.2% male, 14.4% female) and representative of the larger population of AI practitioners (Kaggle 2021) (82.2% male, 16.2% female).

Our participants' age was primarily between 25 and 44 and matches similar studies (Grosse et al. 2023b; Kaggle 2021). To maintain anonymity, we asked for our participants' locations based on dial codes. We received at least one participant from each code area, our sample thus spans 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 samples (Grosse et al. 2023b; Kaggle 2021). In terms of AI background, 5.2% of our participants were trained only, with most (37.3%) having 2-5 years of working experience in AI or ML. Almost as many (35.8%) worked for more than 5 years. 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 larger than 15 (14%). This contrasts previous studies (Kaggle 2021) reporting a quarter of their population in either small or 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 encompasses companies from Africa (2.2%) and Latin America (0.4%). The most frequent industries were healthcare (15.5%), cybersecurity (13.7%), and automotive or a

supplier of automotive (9.2%). Other areas encompassed education (3.3%), arts and entertainment (3,3%), and finance and insurance (4,8%). The remainder were other areas. 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 (Grosse et al. 2023b; Kaggle 2021). AI maturity also coincided with previous samples (Grosse et al. 2023b; Kaggle 2021): 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%).

3 Results

In this short version of the paper, we discuss the specific results of three attacks - backdoors, evasion, and model stealing. For each attack, we revisit the exact threat model from the literature and then discuss whether this commonly used threat model matches the replies from our participants.

3.1 Training-time Attacks

Backdoor attacks rely on perturbing the training data (samples and labels) to affect the resulting model (Ji, Zhang, and Wang 2017; Cinà et al. 2023). Afterward, the backdoor can be used on the trained model (Ji, Zhang, and Wang 2017; Cinà et al. 2023). Consequently, in this section, we first discuss the accessibility of training data and how much of the data can be tampered with, before we discuss the accessibility of the test data and model-reuse.

Access to training data. We asked in how many cases the training data is accessible ($\mathcal{Q}23$). Our participants stated that in 57.1% of cases, the training data was not accessible at all, in 34.4% it was under access control, and only in 8.5%, the data was publicly accessible, e.g. to someone who is neither employee at the company or client. 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 ($\mathcal{Q}23$) and the percentage of training data from public sources ($\mathcal{Q}28$). 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. 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% of the data stemmed from public sources or that they were unsure how much came 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.

Table 2: Comparing assumptions about alterable training data from academia in backdoor (Cinà et al. 2023) papers to our participants' reports. We first state the percent of alterable training data, then the number of backdoor papers with the specific assumptions. Finally, we show the percentage of participants in our sample that stated this amount of data was alterable. The two percentages marked with * were misaligned with our questionnaire and were thus estimated.

Percent training data altered	# Backdoor papers (Cinà et al. 2023)	Our findings
>30%	–	* <30.3%
10-30%	20	* <10.7%
<10%	12	20.4%
∅	—	30.3%

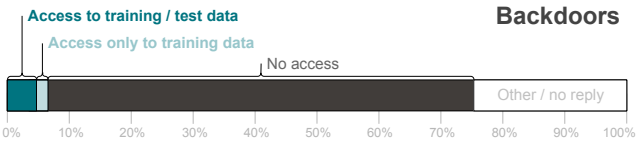

Figure 1: Backdoor threat model in percent of our participants' replies. We report 3rd party access: White denotes incomplete data or an irrelevant threat model (e.g., only accessible test data). Black represents no access, turquoise the backdoor threat model in academic research. Light turquoise denotes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

Percent of data changed. Cinà et al. (2023) surveyed the percentage of training data an attacker altered in 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 ($\mathcal{Q}28$). As before, the heavily studied middle range (10-30%) was the least common in practice.

Access to test data. To submit a 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. While the setting where only training data is available was rare, corresponding works exist. These are poisoning attacks, that do not alter test data or implement backdoors that do not rely on a trigger but instead target a small group of clean samples (Cinà et al. 2023; Geiping et al. 2021). Of 32 papers, 10 rely on this specific threat model (Cinà et al. 2023).

Fine-tuning. Another assumption often made in backdoor attacks is that practitioners rely on existing models and fine-tune these. We combined the information provided by Cinà et al. (2023) about the fine-tuning setting and our partici-

Table 3: Comparing assumptions about used test queries in black-box attacks (Mahmood et al. 2021) and model stealing (Oliynyk, Mayer, and Rauber 2023) papers. We first 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 how many participants stated this amount of queries could be submitted. * indicates that attacks may be possible via transferability.

Possible queries	# Evasion black-box papers (Mahmood et al. 2021)	# Model stealing papers (Oliynyk et al. 2023)	Our findings
$\emptyset$	*	—	36.5%
<100	2	5	15.6%
100-1k	8	9	4.8%
1k-100k	1	16	7.4%
>100k	—	10	1.1%
∞	11	40	15.5%

pants' replies ($\mathcal{Q}21$). Of the 32 backdoor papers, 12 dealt with a fine-tuning setting, e.g., the user 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 a model circumvents accessing the training data.

While there are notable exceptions of papers assuming very small backdoor percentages of less than 3% in vision (Han et al. 2022), object detection (Ma et al. 2022), and point clouds (Xiang et al. 2021), more such work is needed. In addition, we currently do not know which quality checks are implemented for public data, and how an attacker could circumvent these checks.

> **Take away–Training time attacks.** We find evidence that assumptions of backdoor threat models are met in practice. Yet, while data can often not be accessed directly, 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.

3.2 Test-time attacks

Evasion and model stealing target a model at test time (Dalvi et al. 2004; Szegedy et al. 2014; Biggio and Roli 2018; Oliynyk, Mayer, and Rauber 2023). They are thus similar and require submitting 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 ($\mathcal{Q}26$), almost half (45%) of our participants reported that the samples were not accessible at all, almost a quarter (23.2%) that the test data was under access control, and only 5.1% percent that their test data was publicly accessible. The model ($\mathcal{Q}24$) was not accessible for half (49.8%) of our participants, and accessible

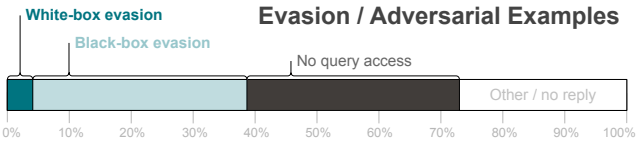

Figure 2: Evasion threat models in percent of our participants' replies. We report 3rd party access. White denotes incomplete data or an irrelevant threat model (e.g., only model accessible). Black represents no access, turquoise white-box and light turquoise black-box evasion threat models.

using authentication for one quarter (25.8%). In 7.7%, the model was publicly available. Model outputs ($\mathcal{Q}25$) were available more readily: Outputs were not accessible in only roughly one quarter (23.6%) of the cases, accessible under access control in one-third (35.4%) of the cases, and freely available in one-fifth (19.2%) of the cases.

As before, test data not accessible directly may be altered when coming from untrusted sources ($\mathcal{Q}27$). We again investigated frequent combinations. The most frequent combination was (14%) no test data from public sources when the test data was not accessible. The second most frequent combination (2.2%) was inaccessible test data where more than 75% stems from public sources. No other combination occurred more than 2% or 5 times in our data.

To understand how many test queries could be submitted when the model was publicly accessible ($\mathcal{Q}33$), we again examined frequent combinations. The most frequent (8.9%) is public model access with unconstrained queries. This was followed by (4.1%) 10-100 queries, (1.8%) 100-1,000, and (1.8%) 100,000-100,000,000 queries. All other combinations occurred less than 5 times. This roughly mirrored the overall distribution of replies: Most participants granted no, few, or unlimited queries.

> **Take away–Test time attacks.** Compared to the training data, the threat surface is larger at test-time but is still small when assuming that business partners are benign. 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 (Biggio and Roli 2018; Dalvi et al. 2004; Gnanasambandam, Sherman, and Chan 2021; Madry et al. 2018). We examined these threat models and visualized our participants' replies in Figure 2. We found that 3rd party access these two features ($\mathcal{Q}24$ and $\mathcal{Q}33$) was rare and only reported by 4.1% of our participants. If we dropped the white-box constraint and permitted the attacker to not consider the model, this percentage increased strongly to 34.6%. As expected, black-box attacks can be carried out more frequently.

We thus focus on black-box attacks (Biggio and Roli 2018; Mahmood et al. 2021; Croce et al. 2021; Garcia et al. 2023) and the needed queries for an attack ($\mathcal{Q}33$). We re-

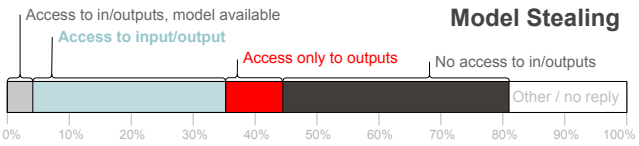

Figure 3: Model stealing threat model in percent of our participants' replies. We describe 3rd party: White denotes incomplete data or an irrelevant threat model (e.g., only test inputs are accessible). Black represents no access, turquoise denotes the academic threat model, gray that the attack is obsolete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

lied on the overview of Mahmood et al. (2021). For the sake of this comparison, we ignore whether attacks are targeted or untargeted and whether hard or soft labels are required, and report the minimal empirical amount of queries documented (Mahmood et al. 2021) in Table 3. Few (2) papers operated in the most frequent setting (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 operates in the range of 1,000-100,000 queries, which is slightly more frequent (7.4%). On the other hand, 15.5% of our participants allowed infinite queries. Access to AI systems in practice was thus all-or-nothing, with few queries or infinitely many. Research, in contrast, focused on the middle range, 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 (Garcia et al. 2023). In addition, our work is a call for transferability studies, when neither model nor data are known, as uttered by Sheatsley et al. (2023). Existing work indicates that these settings are harder than the currently studied settings where the training data is assumed to be known (Alecci et al. 2023). Such a setting (attacking only via test data) was most practical according to our participants.

> **Take away–Evasion.** Current industrial models are vulnerable against white-box evasion in 4.1% of the cases but vulnerable to black-box attacks in a third of the cases. 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. This highlights the need to deepen our understanding of transferability.

Model stealing. Model stealing attacks target the model via test inputs and outputs (Oliynyk, Mayer, and Rauber 2023) ($\mathcal{Q}$24, $\mathcal{Q}$32, and $\mathcal{Q}$33). The goal is to obtain a copy of the target model either in terms of functionality or a direct copy of the weights. We examined this threat model in practice and plotted the corresponding percentages in Figure 3. 44.5% of our participants reported that public access to their model outputs is possible. Most model stealing attacks (Oliynyk, Mayer, and Rauber 2023; Tramèr et al. 2016) require to submit specific queries, decreasing this percentage to 35.3%. In about a seventh of this third, or 4.1%, 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. (2017).

An additional factor in model stealing is, analogous to black-box attacks, the submittable number of queries to the target model ($\mathcal{Q}$33). As before, we compared the number of queries reported by Oliynyk, Mayer, and Rauber (2023) to our sample in Table 3. Most of the 40 papers surveyed required between 100 and 100,000 queries, exactly the numbers our participants reported the least frequently. Only five papers relied on less than 100 queries and thus fell within a larger (15.6%) percentage of our sample. While we so far ignored that not all papers report their sample complexity (Oliynyk, Mayer, and Rauber 2023) and precision of the attack may vary (Oliynyk, Mayer, and Rauber 2023), we further investigate the relationship between the number of queries allowed and model complexity (as approximated by input size, $\mathcal{Q}$29). There is no statistically significant correlation between these two features. 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 occurred 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.

> **Take away—Model stealing.** Model stealing can be carried out in practice. Yet, in a few 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 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 between query number and model complexity.

4 Limitations

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

Sample limitations. Our sample is biased towards the global north, especially Europe, and is limited to English-speaking practitioners. We have not collected cultural or ethnic affiliations, or non-binary gender information and can not exclude that our sample is biased. Although we managed to recruit over 250 participants, we could not find reliable and consistent scientific references to estimate the global target population of industrial AI practitioners. 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 (Kaggle 2021) rather well. As our goal is to identify conceptual mismatches of threat models in the wild compared to research, we find this margin of error is acceptable.

Methodological limitations. We did not review the entire body of AI security work for our analysis. Given that there are several thousand research articles about AI security[1], this endeavor is beyond a single paper. We instead rely on surveys (Cinà et al. 2023; Mahmood et al. 2021; Oliynyk, Mayer, and Rauber 2023; Jegorova et al. 2022; He et al. 2022) 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 (Cinà et al. 2023). The scope of our comparison is thus limited and may be biased. As before, we reason that this overview is sufficient to identify conceptual gaps. As argued, the success of some attacks like model stealing depends on model complexity (Oliynyk, Mayer, and Rauber 2023). As model complexity is not straightforward to measure, we left a detailed analysis for future work. 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 depends on a specific application, which we do not cover but leave for future work.

5 Implications and Future Work

Having discussed the limitations, we are ready to present 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 implications concerning AI regulation and AI security in practice. Where applicable, we also delve into future work for these latter implications.

5.1 Future Work in AI Security

We found several gaps between the academically studied threat models and practical AI usage (Sect. 3). Consequently, most of our implications translate to direct recommendations of previously overlooked aspects. In this section, we attempt to give the big picture by combining our findings for each attack, listing open questions alongside. At the end of the section, we summarize insights that go beyond individual attacks but reply to AI security in general.

Backdoors. Backdoor threat models apply in practice (Sect. 3.1). Further studies should focus on ending the arms-race and deepening our knowledge of defense trade-offs (Cinà et al. 2023). At the same time, current percentages of frequently altered training data are not well aligned with the percentages reported by our practitioners (Sect. 3.1). 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. 3.1), corresponding risks need to be assessed (Hong, Carlini, and Kurakin 2022).

Evasion. We found evidence of the applicability of (black-box) evasion threat models (Sect. 3.2), and recommend further study to end the arms-race (Croce et al. 2021; Tramer et al. 2020). Still, more emphasis should be put on studying attacks that succeed without knowledge of the exact data and model specifics (Sect. 3.2). This is aligned with previous observations that more work is required on transferability. More precisely, and as stated by Sheatsley et al. (2023), we should not only study model-to-model transferability, but also transferability across different datasets. If queries are allowed, the number of queries should be minimized, ideally to less than 100, to reflect practice (Sect. 3.2).

Model stealing. We found evidence of the applicability of model stealing threat models (Sect. 3.2). Future work should address the corresponding arms-race (Oliynyk, Mayer, and Rauber 2023). We further found a mismatch of used queries in model stealing and a mismatch for the attacker's capabilities overall (Sect. 3.2). Consequently, we recommend reducing used queries, and not relying on currently reported high amounts of queries, similar to evasion. Furthermore, observing only outputs is a valid, currently scarcely studied threat model. It may thus be beneficial to understand the limitations of retrieving information only by observing outputs (Papernot et al. 2017). 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 or model extraction (Jegorova et al. 2022).

5.2 Practical Implications

Our research has however implications beyond AI security research, relating to both regulations and required knowledge about AI security in practice, which we discuss now.

Regulatory and societal implications. Assessing the true vulnerability of AI in practice has implications for regulation, as current proposals like the EU AI Act demand that training data be secure. For example, knowing few models are accessible to 3rd parties in practice (Sect. 3) may imply that similar requirements be a possibility to enhance security. A complete analysis of used access schemes is however most likely related to use-case, industry area, and other factors, and thus left for future work. Beyond regulation, assessing vulnerabilities helps to manage the risk of potential security incidents. Using our threat models (Sect. 3), 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. We find that all 3 attacks studied within the framework of AI security are theoretically possible in practice. The percentage of our 271 participants reporting the exact required access means to conduct current attacks is however small, potentially contributing to an explanation of formerly found low percentages of security in-

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

cidents (Grosse et al. 2023b). Analogously, we need to understand the limitations of knowing data at all in practice. In tasks such as malware detection, feature encodings are secret, limiting the attacker (Biggio and Roli 2018). 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. 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 AI vulnerabilities. Finally, we have not studied AI vulnerabilities from the defender's perspective, a valuable endeavour.

6 Related Work

Several contributions criticize existing scientific threat models in AI security (Gilmer et al. 2018; Evtimov et al. 2020; Apruzzese et al. 2022; Cinà et al. 2023). To the best of our knowledge, none of these works provides an overall picture of the research gap in AI security in practice.

Few works collect loosely similar information to our work, including the Kaggle (2021) annual report about ML and data science, which provides information on, for example, the algorithms used in practice. Furthermore, Nahar et al. (2022) investigated the origin of the used data, yet in a small qualitative sample. They also document the influence data engineers have on data requirements (Nahar et al. 2022). Dilhara, Ketkar, and Dig (2021) studied the usage of libraries in ML-based code, but based their study on available repositories, not industry applications. Renieris, Kiron, and Mills (2023) examined the practical usage of third-party tools and found that almost three-quarters use such tools. The same authors show that the same tools may cause AI failures. Finally, Mink et al. (2023) investigate the number of deployed AI security mitigations in practice.

Previous works reported low (Boenisch et al. 2021) to medium (Grosse et al. 2023b) 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 systems in practice.

7 Conclusion

We took a significant step towards more practical AI security research. We surveyed common threat model properties in practice, including data and model access, the number of queries, and data from public sources. We then matched this information to three 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 and paved the way towards 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. In general, current threat models are too generous about training data access or test time queries. The prevailing assumptions on attacker knowledge in, for example, model stealing, and backdoors are too generous. Our paper is thus a call for action to study more practical threat models in AI security research.

Acknowledgement

We would like to thank all participants and are grateful for the support of Hyrum Anderson, Marielle Dado, Daryan Dehghanpisheh, Nikki Hogg, Ritesh Sharma, Aryan Trip, Karn Wong, Alla Zhdan, and the MLOps community.

References

Alecci, M.; Conti, M.; Marchiori, F.; Martinelli, L.; and Pajola, L. 2023. Your Attack Is Too DUMB: Formalizing Attacker Scenarios for Adversarial Transferability. In *RAID*, 315–329.

Apruzzese, G.; Anderson, H.; Dambra, S.; Freeman, D.; Pierazzi, F.; and Roundy, K. 2022. Position:"Real Attackers Don't Compute Gradients": Bridging the Gap Between Adversarial ML Research and Practice. In *IEEE Conf. on Secure and Trustworthy Machine Learning*. IEEE.

Barreno, M.; Nelson, B.; Sears, R.; Joseph, A. D.; and Tygar, J. D. 2006. Can machine learning be secure? In *CCS*, 16–25.

Bieringer, L.; Grosse, K.; Backes, M.; and Krombholz, K. 2022. Mental Models of Adversarial Machine Learning. In *SOUPS*, 97–116.

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

Boenisch, F.; Battis, V.; Buchmann, N.; and Poikela, M. 2021. "I Never Thought About Securing My Machine Learning Systems": A Study of Security and Privacy Awareness of Machine Learning Practitioners. In *Mensch und Computer*, 520–546.

Chakraborty, A.; Alam, M.; Dey, V.; Chattopadhyay, A.; and Mukhopadhyay, D. 2021. A survey on adversarial attacks and defences. *CAAI Transactions on Intelligence Technology*, 6(1): 25–45.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv:1712.05526*.

Cinà, A. E.; Grosse, K.; Demontis, A.; Vascon, S.; Zellinger, W.; Moser, B. A.; Oprea, A.; Biggio, B.; Pelillo, M.; and Roli, F. 2023. Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning. *ACM Comput. Surv.*, 55(13s).

Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; Chiang, M.; Mittal, P.; and Hein, M. 2021. RobustBench: a standardized adversarial robustness benchmark. In *NeurIPS Datasets and Benchmarks Track*.

Dalvi, N.; Domingos, P.; Mausam; Sanghai, S.; and Verma, D. 2004. Adversarial Classification. In *KDD*, 99–108.

Dilhara, M.; Ketkar, A.; and Dig, D. 2021. Understanding Software-2.0: A Study of Machine Learning library usage and evolution. *ACM Transactions on Software Eng. and Methodology (TOSEM)*, 30(4): 1–42.

Evtimov, I.; Cui, W.; Kamar, E.; Kiciman, E.; Kohno, T.; and Li, J. 2020. Security and Machine Learning in the Real World. *arXiv:2007.07205*.

Garcia, W.; Chen, P.-Y.; Clouse, H. S.; Jha, S.; and Butler, K. R. 2023. Less is More: Dimension Reduction Finds On-Manifold Adversarial Examples in Hard-Label Attacks. In *IEEE Conf. on Secure and Trustworthy Machine Learning*, 254–270.

Geiping, J.; Fowl, L. H.; Huang, W. R.; Czaja, W.; Taylor, G.; Moeller, M.; and Goldstein, T. 2021. Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching. In *ICLR 2021*.

Gilmer, J.; Adams, R. P.; Goodfellow, I.; Andersen, D.; and Dahl, G. E. 2018. Motivating the rules of the game for adversarial example research. *arXiv:1807.06732*.

Gnanasambandam, A.; Sherman, A. M.; and Chan, S. H. 2021. Optical adversarial attack. In *ICCV*, 92–101.

Grosse, K.; Bieringer, L.; Besold, T. R.; and Alahi, A. 2023a. Towards more Practical Threat Models in Artificial Intelligence Security. arXiv:2311.09994.

Grosse, K.; Bieringer, L.; Besold, T. R.; Biggio, B.; and Krombholz, K. 2023b. Machine learning security in industry: A quantitative survey. *IEEE Transactions on Inf. Forensics and Security*, 1749–1762.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the ML model supply chain. *arXiv:1708.06733*.

Han, X.; Xu, G.; Zhou, Y.; Yang, X.; Li, J.; and Zhang, T. 2022. Physical Backdoor Attacks to Lane Detection Systems in Autonomous Driving. In *ACM Int. Conf. on Multimedia*, 2957–2968.

Harris, P. A.; Taylor, R.; Minor, B. L.; Elliott, V.; Fernandez, M.; O'Neal, L.; McLeod, L.; Delacqua, G.; Delacqua, F.; Kirby, J.; et al. 2019. The REDCap consortium: building an Int. community of software platform partners. *Journal of biomedical informatics*, 95: 103208.

He, X.; Li, Z.; Xu, W.; Cornelius, C.; and Zhang, Y. 2022. Membership-Doctor: Comprehensive Assessment of Membership Inference Against Machine Learning Models. *arXiv:2208.10445*.

Hong, S.; Carlini, N.; and Kurakin, A. 2022. Handcrafted backdoors in deep neural networks. *NeurIPS*, 35: 8068–8080.

Jegorova, M.; Kaul, C.; Mayor, C.; O'Neil, A. Q.; Weir, A.; Murray-Smith, R.; and Tsaftaris, S. A. 2022. Survey: Leakage and privacy at inference time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Ji, Y.; Zhang, X.; and Wang, T. 2017. Backdoor attacks against learning systems. In *IEEE CNS*, 1–9.

Kaggle. 2021. State of Machine Learning and Data Science. https://www.kaggle.com/kaggle-survey-2021.

Ma, H.; Li, Y.; Gao, Y.; Abuadbba, A.; Zhang, Z.; Fu, A.; Kim, H.; Al-Sarawi, S. F.; Surya, N.; and Abbott, D. 2022. Dangerous Cloaking: Natural Trigger based Backdoor Attacks on Object Detectors in the Physical World. *arXiv:2201.08619*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.

Mahmood, K.; Mahmood, R.; Rathbun, E.; and van Dijk, M. 2021. Back in black: A comparative evaluation of recent state-of-the-art black-box attacks. *IEEE Access*, 10: 998–1019.

Mink, J.; Kaur, H.; Schmüser, J.; Fahl, S.; and Acar, Y. 2023. "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*.

Nahar, N.; Zhou, S.; Lewis, G.; and Kästner, C. 2022. Collaboration Challenges in Building ML-Enabled Systems: Communication, Documentation, Engineering, and Process. *Organization*, 1(2): 3.

Oh, S. J.; Schiele, B.; and Fritz, M. 2019. Towards reverse-engineering black-box neural networks. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 121–144.

Oliynyk, D.; Mayer, R.; and Rauber, A. 2023. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys*.

Papernot, N.; McDaniel, P.; and Goodfellow, I. 2016. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv:1605.07277*.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical black-box attacks against machine learning. In *Asia CCS*, 506–519.

Renieris, E. M.; Kiron, D.; and Mills, S. 2023. Building Robust RAI Programs as Third-Party AI Tools Proliferate. *MIT Sloan Management Review*.

Sheatsley, R.; Hoak, B.; Pauley, E.; and McDaniel, P. 2023. The Space of Adversarial Strategies. In *USENIX Security*, 3745–3761.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR*.

Tramer, F.; Carlini, N.; Brendel, W.; and Madry, A. 2020. On adaptive attacks to adversarial example defenses. *NeurIPS*, 1633–1645.

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing Machine Learning Models via Prediction APIs. In *USENIX*, 601–618.

Xiang, Z.; Miller, D. J.; Chen, S.; Li, X.; and Kesidis, G. 2021. A backdoor attack against 3d point cloud classifiers. In *ICCV*, 7597–7607.