When Your AI Becomes a Target: AI Security Incidents and Best Practices

Kathrin Grosse[1], **Lukas Bieringer**[2], **Tarek R. Besold**[3], **Battista Biggio**[4], **Alexandre Alahi**[1]

[1]École Polytechnique Fédérale de Lausanne, Switzerland
[2]QuantPi, Germany
[3]Technical University Eindhoven, The Netherlands
[4]University of Cagliari, Italy
kathrin.grosse@epfl.ch

Abstract

In contrast to vast academic efforts to study AI security, few real-world reports of AI security incidents exist. Released incidents prevent a thorough investigation of the attackers' motives, as crucial information about the company and AI application is missing. As a consequence, it often remains unknown how to avoid incidents. We tackle this gap and combine previous reports with freshly collected incidents to a small database of **32** AI security incidents. We analyze the attackers' target and goal, influencing factors, causes, and mitigations. Many incidents stem from non-compliance with best practices in security and privacy-enhancing technologies. In the case of direct AI attacks, access control may provide some mitigation, but there is little scientific work on best practices. Our paper is thus a call for action to address these gaps.

Introduction

The reliability of Artificial Intelligence (AI) is brittle in adversarial settings (Dalvi et al. 2004; Barreno et al. 2006; Tramèr et al. 2016; Chen et al. 2019; Chakraborty et al. 2021; Cinà et al. 2023; Oliynyk, Mayer, and Rauber 2023). Examples include training time attacks like poisoning (Barreno et al. 2006; Cinà et al. 2023), and test time attacks like evasion (Szegedy et al. 2014; Chakraborty et al. 2021), both decreasing the model's performance. In addition, a model can be copied without consent at test time (Tramèr et al. 2016; Oliynyk, Mayer, and Rauber 2023). Although the above attacks focus on classification, similar exist for reinforcement learning (Dalvi et al. 2004; Chen et al. 2019) or data mining and data analysis (Rubinstein et al. 2009; Chen et al. 2017). Recent efforts systematize this AI security knowledge for usage in practice[1].

Contrasting this academic interest, few reports exist of real-world AI security incidents, or in other words, real-world incidents of the attacks described above. Recent examples include data leakage from large-language models[2], failed chat-bots learning offensive language from their users[2] or instances of decreased quality of search engines[2]. These and similar incidents are included in databases[3](McGregor 2021), which however also contain cases of bias[4] or hardware faults[4]. Many incidents consist of media descriptions – while such information is undeniably valuable, it does not uniquely focus on security and limits a deep understanding of the failures: For example, we may not know all the details about the deployed system, how mature and well-tested it was, the expertise and concern of the company, and what exact component was affected by the attack.

Real-world AI incidents have been also reported as descriptions by study participants (Bieringer et al. 2022) and in surveys (Grosse et al. 2023b), where about 5% of AI practitioners had experienced AI-specific attacks on their AI systems. In addition, Grosse et al. reported no correlations with company size, organizational area, company size, and estimated likelihood of becoming the victim of an attack.

Existing works thus rely on publicly available data, do not focus on AI security (Durso et al. 2022; Pittaras and McGregor 2023; McGregor 2021) or lack an in-depth analysis of incidents (Bieringer et al. 2022; Grosse et al. 2023b). In this paper, we analyze previously reported failures (Bieringer et al. 2022; Grosse et al. 2023b) and ask an additional 271 participants whether they have encountered incidents. Our final dataset encompasses **32** AI failure descriptions in total, almost all with information like company size, and expected likelihood of the attack. Our findings are three-fold. First, only company size—not the time the model has been in production—correlates with more frequent incidents. Second, the majority of incidents target infrastructure or data and compromise the underlying system's confidentiality or integrity. Thirdly, best practices for security and privacy should be followed when developing AI. A focus should be access control, which may also mitigate some attacks directly on the AI. More work is needed to understand AI security vulnerabilities, their development over time, influences thereon, and corresponding mitigations.

Methodology

Our dataset combines reported incidents and newly collected incident reports. More precisely, we considered one incident described by Bieringer et al. (2022) and used the incidents from Grosse et al. (2023b). In this section, we describe our

[1]https://airisk.io, https://attack.mitre.org

[2]Incidents 352, 6 and 88 in https://incidentdatabase.ai/

[3]https://atlas.mitre.org/studies or https://avidml.org

[4]Incidents 101 and 31 in https://incidentdatabase.ai/

questionnaire and recruiting strategy to collect new incidents and our resulting sample. A Full description of questionnaire design, recruiting, and sample comparison can be found in our orthogonal paper (Grosse et al. 2023a). We first review and define basic terminology and concepts for our analysis.

Terminology and Definitions

Security studies the effect an attacker can have on an existing system or program. We, thus, first define three goals an attacker can have, before we review different attacks and areas like AI security, cyber security, or safety.

Confidentiality refers to sensitive information being protected against illegitimate access. A subaspect is the access to, e.g., health data, also called personally identifiable information (PII), leading to a **privacy** incident.

Integrity refers to the consistency, accuracy, and trustworthiness of data during their lifecycle. In other words, sensitive data or code should not be alterable.

Finally, **availability** refers to data being available to the legitimate user when needed. This includes data, hardware, technical infrastructure, and systems.

Based on these security concepts, we can now review AI security and contrast it with both cyber security and safety.

Poisoning attacks alter the training data, samples, or labels, to decrease the overall performance of the classifier (Barreno et al. 2006; Cinà et al. 2023). They, thus, target the availability of the AI model.

Model stealing (M.S.) copies the model without the owner's consent by submitting tailored inputs to a model (Tramèr et al. 2016; Oliynyk, Mayer, and Rauber 2023). M.S. harms confidentiality, as the model is leaked.

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), thus harming the model's integrity.

AI security in this paper encompasses all the above attacks: poisoning, evasion, and model stealing.

Cybersecurity includes any breach, attack, or circumvention that is not covered by AI security.

We furthermore borrow the term **safety** from systems analysis. It denotes, in contrast to a security incident caused by an attacker, a failure that is benign and not caused by an adversary (Muller, Young, and Vogt 2007).

Questionnaire and Recruiting

To collect data about AI incidents, we also queried some demographic information to compare to other samples, containing variables such as age, gender, geographic location, company size, industry area, team size, and whether and how long the AI system was in production. Afterward, similar to Grosse et al. (2023b), we inquired about encountered incidents and asked our participants for a description of the experienced incidents and the number of occurrences (1,2,3,4, > 4). We further asked our participants to estimate the likelihood of experiencing a security incident, and whether and which parts of their AI system were publicly accessible. Due to the sensitive nature of some of our questions—experienced attacks—we opted for an anonymous, unpaid questionnaire containing only multiple-choice questions. An exception was the written description of the incident to not constrain sharing details. All fields could be left blank to allow for confidentiality.

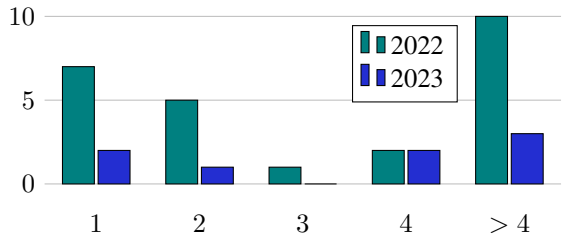

Figure 1: Occurence of individual AI incidents.

After obtaining approval from our institution's ethical review board, we implemented the questionnaire using RedCap (Harris et al. 2019) and conducted several 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 these the most likely to be aware of AI security incidents.

Sample and Data Analysis

In contrast to the 15% incidents reported by Grosse et al. (2023b), our sample of 271 participants reported only 3% attacks on AI workflows. A possible reason could be that Grosse et al.'s questionnaire focused on AI security, while we advertised our study as a general AI questionnaire to avoid priming. The time of recruiting was roughly the same, 3 and 2.5 months, respectively. Otherwise, the demographics in terms of age, gender, AI education, education, and company size matched both the previous sample (Grosse et al. 2023b) and the larger distribution (Kaggle 2021).

For further analysis, the first and fourth authors categorized each incident according to the area (AI security, cybersecurity, etc) and the likely attacker's goal (confidentiality, integrity, and availability). All incidents and their assignment are in Table 1. The categorizations were suggested by one author, and then jointly refined by both authors.

Before we analyze our data, we describe the basic statistics of the 32 incidents. While there were 23 incidents in the 2022 sample, of which 6 (26%) were AI security-related, our sample contained 8 incidents, of which 2 (25%) were AI security-related. Our sample contained a further 3 (37.5%) cyber security incidents. In contrast, the 2022 sample comprised 6 (26%) cybersecurity, and 4 (17.4%) privacy incidents. In both samples, there is a high number of incidents for which we could not assign a category, as the provided information was insufficient. In the 2022 data, 6 (26.1%) fell in this category, and in our data 3 (37.5%).

Experimental Results

We first discuss possible influences on AI incidents like company size, then give an overview of the targeted components, before we discuss the incidents in detail.

Description of incident		Target			Time		Goal		
		D	M	Inf	Tr	T	C	I	A
"That answer is just a guess, I know people run crypto mining on our infra through our APIs"	cysec			◇				▽	
"We got hit by crypto-miners pretty hard trying to abuse our free tier or use stolen credit cards to use us."	cysec			◇					(▽)
"exposure of confidential research data to cloud services"	cysec	◇					▽		
"Server was compromised"	cysec			◇			▽		
"Universal bypass"	cysec						▽		
"incorrect data access"	cysec	◇					▽		
"phishing"	cysec			◇					
"Using manual overrides without prior authorization"	cysec	◇		◇					
"Botnet communication"	cysec			◇					
"A man in the middle attack between two workflows when data was being transferred through the public internet"	cysec	◇							
"Real world data is fed back into our models and also provided to our customers to enable them to learn themselves. Incorrect use of temporary storage on AWS lambda led to a potential leak of raw inference input data (later to be scrubbed for training data), that could potentially contain PII to the wrong customer. Luckily, this exploit was discovered before a real customer could interact with the feature."	privacy	◇		◇	●	●	▽		
"patient data should not be shared."	privacy	◇					▽		
"data privacy"	privacy	◇					▽		
"Anonymizing medical data from patients before using them in training"	privacy	◇					▽		
"Acquiring the Data for training AI Systems"	unclear		◇				▽		
"ML systems being retrained to provide false outputs"	poisoning	◇	◇		●			▽	▽
"1 type, but happened multiple times. Client / partner employees tasked with labeling training data feel threatened by automation, and either stall or sabotage the labeling effort, harming the models. This is not an "outside" threat, but hard to protect against once it happens."	poisoning	◇			●		▽		
"DDOS attacks and probing rule-based solutions"	M.S., DDoS		◇				▽	▽	
"people hitting the endpoint trying to reverse engineer the way we got results"	M.S.		◇				▽		
"Autonomous vehicle image recognition errors leading to dangerous path planning. Not an 'intentional' circumvention, but it bypassed automated safety checks nonetheless."	evasion*	◇				●		▽	
"object detection and semantic classification of video sensors on cars: a speed-sign as advertisement put onto a driving-school car was interpreted by the camera as a valid street sign, sending that information to the software driving function (which was expecting Autobahn speed but saw a 30km/h speed-sign)"	evasion*	◇				●		▽	
"What we found is [...] common criminals doing semi-automated fraud using gaps in the AI or the processes, but they probably don't know what AML, like adversarial machine learning is and that they are doing that. So we have seen plenty of cases are intentional circumventions, we haven't quite seen like systematic scientific approaches to crime"	evasion	◇				●		▽	
"users spam to optimize their strategy for job search"	evasion	◇				●		▽	

Incidents providing insufficient information for analysis (9):

"Complete redesign", "I refer to Meta and Amazon who do this everyday with my personal workflows !!!!"
"Brute Force Attacks", "Brute force attack","We have not had many severe circumeventions. We place user privacy at highest priority"
"Unable to furnish the details publicly", "No details", "Can't disclose", " "

Table 1: All incidents with our systematization. We distinguish as targets within the incident **D**ata, **M**odel and **Inf**rastructure, the time of the attack (**Tr**aining or **T**est time), and the goal of the attacker: **C**onfidentiality, **I**ntegrity, and **A**vailability. M.S. translates to model stealing attacks. All poisoning, evasion, and model stealing are subgroups of AI security. The abbreviation DDoS refers to distributed denial of service, PII to personally identifiable information, and AML for adversarial machine learning, or, in other words, machine learning security.

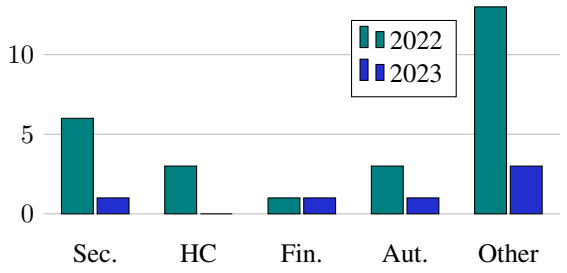

Figure 2: Number of incidents per industry area. Sec. is security, HC healthcare, Fin. is finances, Aut. automotive and other none of the previous.

Influences on AI Incidents

We analyze possible factors influencing AI incidents. We first examine the industry area, also visualized in Figure 2. The largest represented industry was security with 7 (22%) incidents, followed by automotive with 4 (12.5%), healthcare with 3 (9.4%), and finance and insurance with 2 (6%). These industries were also prevalent in the overall sample (e.g., without reported incidents). However, the full sample encompassed 57 healthcare workers and 50 security practitioners, indicating that AI security incidents occurred or are reported more frequently in some industries (like security).

Independent of the application area, one may assume that larger companies or companies whose AI products are longer on the market are more likely to experience or fear AI security-related incidents. We tested this hypothesis using Spearman correlation across the corresponding features. The number of incidents and the estimated likelihood of becoming a victim correlate positively (0.49). In other words, the more incidents a practitioner witnessed, the more concern they expressed (or vice versa). We also found a correlation (0.29) between the number of incidents and the company size. Rephrasing, more incidents were experienced when the company was larger. In all other cases, a linear correlation was absent, for example between the estimated likelihood of experiencing an attack and the company size (-0.02) or whether AI was used in a product and how long (0.05). Finally, there is no correlation between the number of incidents and whether and how long AI was used in products (0.09). While the latter could mean that also systems fresh on the market are targeted, or that incidents depend on the application area. However, more work is needed to determine an in-depth understanding of these findings.

Concluding, in our sample, company size is the feature with the strongest correlation to the number of incidents.

Vulnerabilities within AI

From our incidents, we learn that the target of attacks for AI incidents is most often the data, followed by the infrastructure, and then by the model. In terms of time, most attacks are independent of timing within the ML pipeline, e.g. were not attributable to training, test, or evaluation. The second most frequent timing was during test time, then training time. Concerning the attacker's goal, we found that confidentiality was the most frequent goal within our incidents, followed by integrity and availability.

		Target			Time		Goal		
#		D	M	Inf	Tr	T	C	I	A
6	cysec	◇ *		◇ *			▽ *		
5	cysec			◇				▽ *	▽ *
3	privacy	◇					▽		
1	privacy	◇		◇	●	●	▽		
1	poisoning	◇	◇		●			▽	▽
1	poisoning	◇			●		▽		
2	M.S.		◇				▽	▽ *	
4	evasion	◇				●		▽	

Table 2: A condensed overview of our incidents, grouped according to target, timing, and goal of the incident or attack. ◇, ●, and ▽ denote all incidents share the target, time or goal; ◇ * and ▽ * denote that some incidents do.

Causes and Fixes

We also analyze our incidents to derive possible mitigations. There are 24 (75%) incidents providing enough information for analysis. We attempt to deduce the incident's cause and summarize grouped incidents in Table 2.

Cybersecurity. The first incident group within cybersecurity contained six incidents with varying targets and goals. The shared characteristic was that often, data was the target, and the goal was a confidentiality breach. In some cases, the infrastructure was the target. The timing within the ML pipeline was not clear. Participants stated "Universal bypass" or "Using manual overrides without prior authorization", or "A man in the middle attack [...] when data was being transferred [...]". In addition to proper access control to prevent access to confidential resources, best security practices could prevent man-in-the-middle attacks.

On the other hand, the second group shared the characteristic that the infrastructure was affected. The common attacker's goal was integrity or availability. As before, there is no time-wise relation to the ML application. For example, participants noted "phishing," that a "server was compromised," or "people run[ning] crypto mining on our infra through our APIs". Analogous to before, access control and security best practices can help protect the infrastructure against attacks like resource theft or malware infections.

Privacy. Three incidents were described only roughly, but a relation to privacy was evident. Descriptions included "patient data should not be shared" or simply "data privacy." The relation to AI (beyond being data) was vague, and might not be given at all. In other words, these attacks most likely referred to data, not model privacy, and the attacker's goal would be a confidentiality breach. In addition, one participant wrote: "Incorrect use of temporary storage on AWS lambda led to a potential leak of raw inference [...] potentially containing PII [...], " indicating that infrastructure was involved. Incidents may thus be caused not only by an attacker but also by misconfigurations ("incorrect use"), highlighting again the need to adhere to security guidelines and

best practices. In addition, using privacy-enhancing technologies may be beneficial when dealing with PII.

AI security. The remaining eight incidents were related to AI security. We discuss the three attacks, poisoning, model stealing, and evasion, separately.

Poisoning. There were two poisoning incidents. The first incident's participant remained vague: "ML systems being retrained to provide false outputs", indicating poisoning attacks that potentially affected the integrity or availability of the model. It remained unclear what caused the incident, making it hard to derive a recommendation. The second participant provided more details: "client/partner employees [...] feel threatened by automation, and either stall or sabotage the labeling effort, harming the models." In addition to the poisoning attack, the participant revealed an ethical issue related to the threat of losing one's job due to AI.

Model Stealing. There were two model stealing incidents. The participants described that "people hitting the endpoint trying to reverse engineer the way we got results", or "probing rule-based solutions". In both cases, submitted inputs and probably the corresponding observed outputs were used to reconstruct (some functionality of) the model. Protection against such an attack can only be provided by access control to inputs, outputs, and the model itself.

Evasion. Somewhat related, but with a slightly different goal of the attacker, one participant wrote that "users spam to optimize their strategy for job search". In this case, users submitted queries to carry out an evasion attack. Another participant described evasion on a high level as "intentional circumventions". Here, it may also help to implement access control to protect the model. There were two similar evasion attacks, one described as "a speed-sign as advertisement put onto a driving-school car was interpreted by the camera as a valid street sign." Both incidents were, according to the participants "not an 'intentional' circumvention, but they bypassed automated safety checks." Intriguingly, the two were half of the four automotive industry cases, with the other having no description at all. Other industries' incidents were, in contrast, more diverse.

While evasion can be alleviated partially by using access control, there is no AI-based solution, as defending evasion is an open question (Chakraborty et al. 2021). Models can however be hardened or tested with evasive samples to assess the risk (Chakraborty et al. 2021).

Conclusion In many cases, incidents may have been prevented by adhering to security best practices and enhancing privacy. This holds even for AI security, where access control may alleviate threats. Some issues, like evasion attacks, are currently undefended, making it difficult to define best practices. Furthermore, several incidents are not strictly security, but bugs or safety issues.

Limitations

The data analyzed in this paper rely on an English questionnaire and—despite efforts to recruit globally—is biased towards the global north and the Western world. Our sample is, with 32 incidents, small. However, obtaining our incidents required recruiting more than 400 participants, of which very few experienced incidents. The information about the incidents is limited, and many (25%) incidents were unclear. Despite these limitations, we believe that sharing existing AI security incidents for practical AI security research is a worthwhile endeavor.

Best Practices and Future Work

Having discussed the limitations, we describe the three-fold implications alongside future work. The first two implications focus on best practices for security and privacy and AI security. For AI security best practices, more knowledge about AI security in practice is required, as discussed last.

Security and Privacy Best Practices. The first and foremost implication of our work is that AI applications, as all computer programs or products, should adhere to best practices in security and privacy. Security standards, in particular concerning access control, should always be implemented. This is especially relevant as in many incidents, the infrastructure, not the ML model, was targeted. The most frequent target is the data, outlining the need for proper access control. When using PII within a project, privacy-enhancing technologies should be used to minimize leaks, in addition to the former recommendations.

AI Security Best Practices. Our incidents show the importance of controlling access to the model, its queries, and outputs. How (much) the model can be defended without access control is subject to future work (Oliynyk, Mayer, and Rauber 2023; Chen et al. 2019). Two attacks in our sample, evasion, and model stealing, depend on a solution to the ongoing arms-race (Oliynyk, Mayer, and Rauber 2023; Chen et al. 2019). To improve model safety, we suggest tests based on, for example, evasion attacks (Chen et al. 2019). The exact definition and nature of the required tests remain future work. Beyond our recommendations, we need more work to develop best practices for AI security and their effect on AI practitioners. While there are suggestions from industry[5], there is little scientific work on the topic so far.

Effect of Best Practices Using the above countermeasures, all reported cybersecurity attacks are avoidable, three of four privacy attacks, and one of two poisoning attacks. In the cases where these are not avoidable, it is usually due to unclarities of the incident description. In contrast, for model stealing and evasion, we can not determine whether attacks are avoidable as deployment and application details are unknown. Overall, this yields 11 unclear cases (34.4%), 15 avoidable (46.9%) incidents, and the remaining six incidents requiring more research (18.8%).

Researching AI Security in Practice. Our results show the need for an in-depth understanding of AI security in practice. Despite initial works (Bieringer et al. 2022; Grosse et al. 2023b; McGregor 2021; Durso et al. 2022), more AI security incidents are needed to understand causes and countermeasures. While our contribution is, with 32 analyzed incidents, a first step towards this goal, more work is needed to monitor attacks over time, understand how individual industries are affected, and improve the incident description

[5]https://github.com/Azure/AI-Security-Risk-Assessment/blob/main/AI_Risk_Assessment_v4.1.4.pdf

quality. Many participants are also reluctant to share details. It may be beneficial to provide an anonymous environment to submit incidents or provide an environment in which AI security incidents can be shared anonymously.

Related Work

We share part of the data with Grosse et al. (2023b). However, they do not focus on the analysis of incidents themselves, the target or goal of the attacker, or derive insights to prevent incidents. On the other hand, we confirm their participants' reports that data is crucial for the security of AI.

Loosely related to our work are taxonomy proposals to collect AI incidents for root cause analysis (Durso et al. 2022; Pittaras and McGregor 2023). Furthermore, best practices in safety and governance have been suggested by Schuett et al. (2023), with a focus on general artificial intelligence. To the best of our knowledge, there are no general AI systems yet. Finally, the AI incident database has been analyzed, but from the perspective of ethical and privacy incidents (Wei and Zhou 2022). The same authors find similar to our sample that frequent industries affected by AI incidents are, among others, autonomous driving, healthcare, and finance. Security is, however, not listed as a specific industry, but factors like authentification are named explicitly.

Conclusion

We have analyzed the first research dataset of AI incidents according to described attack, attacker's target, and goal. Only company size, not the time the model is in production correlated with more incidents. Most incidents target data or infrastructure and harm the confidentiality or integrity of the underlying system. We conclude that in AI development, we should adhere to best practices in security and privacy. A focus should be access control, which may mitigate attacks directly on the AI. More work is needed to understand AI security vulnerabilities, their development over time, influences thereon, and corresponding mitigations. Our paper is thus a call for more research on AI-security best practices.

Acknowledgements

We would like to thank all participants.

References

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.

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, T.; Liu, J.; Xiang, Y.; Niu, W.; Tong, E.; and Han, Z. 2019. Adversarial attack and defense in reinforcement learning-from AI security view. *Springer Cybersecurity*, 2: 1–22.

Chen, Y.; Nadji, Y.; Kountouras, A.; Monrose, F.; Perdisci, R.; Antonakakis, M.; and Vasiloglou, N. 2017. Practical attacks against graph-based clustering. In *CCS*, 1125–1142.

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 Computing Surveys*, 1–39.

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

Durso, F.; Raunak, M.; Kuhn, R.; and Kacker, R. 2022. Analyzing Failures in Artificial Intelligent Learning Systems (FAILS). In *IEEE Software Technology Conf. (STC)*, 7–8.

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 Information Forensics and Security*, 18: 1749–1762.

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 international community of software platform partners. *Elsevier Journal of biomedical informatics*, 95: 103208.

Kaggle. 2021. State of Machine Learning and Data Science. https://www.kaggle.com/kaggle-survey-2021. Accessed: 2023-12-04.

McGregor, S. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. In *AAAI*, 17, 15458–15463.

Muller, P. J.; Young, S. E.; and Vogt, M. N. 2007. Personal rapid transit safety and security on university campus. *Transportation research record*, 2006(1): 95–103.

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*.

Pittaras, N.; and McGregor, S. 2023. A taxonomic system for failure cause analysis of open source AI incidents. *SafeAI@AAAI*.

Rubinstein, B. I.; Nelson, B.; Huang, L.; Joseph, A. D.; Lau, S.-h.; Rao, S.; Taft, N.; and Tygar, J. 2009. Stealthy poisoning attacks on PCA-based anomaly detectors. *ACM SIGMETRICS Performance Evaluation Review*, 37(2): 73–74.

Schuett, J.; Dreksler, N.; Anderljung, M.; McCaffary, D.; Heim, L.; Bluemke, E.; and Garfinkel, B. 2023. Towards best practices in AGI safety and governance: A survey of expert opinion. *arXiv:2305.07153*.

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

Tramèr, F.; Zhang, F.; Juels, A.; Reiter, M. K.; and Ristenpart, T. 2016. Stealing machine learning models via prediction {APIs}. In *USENIX Security*, 601–618.

Wei, M.; and Zhou, Z. 2022. Ai ethics issues in real world: Evidence from ai incident database. *arXiv:2206.07635*.