"*Why do so?*" — A practical perspective on machine learning security

Kathrin Grosse [*1] **Lukas Bieringer** [*2] **Tarek R. Besold** [3] **Battista Biggio** [1 4] **Katharina Krombholz** [5]

Abstract

Despite the large body of academic work on machine learning security, little is known about the occurrence of attacks on machine learning systems in the wild. In this paper, we analyze the answers of 139 industrial practitioners to a quantitative questionnaire about attack occurrence and concern. We find evidence for circumventions of AI systems in practice, although these are not the sole concern of our practitioners, as their reasoning on relevance and irrelevance of machine learning attacks is complex. Our work pave the way for more research about adversarial machine learning in practice, but yield also insights for machine learning regulation and auditing.

1. Introduction

A large body of academic work focused on adversarial machine learning (AML), or the study of how to attack and defend machine learning (Barreno et al., 2006; Biggio & Roli, 2018; Chen et al., 2017; Cinà et al., 2022; Dalvi et al., 2004; Gu et al., 2017; Ji et al., 2017; Oh et al., 2019; Papernot et al., 2016; Szegedy et al., 2014; Tramèr et al., 2016). However, few work go beyond artificial, simple settings and study such attacks on a traffic sign recognition system (Woitschek & Schneider, 2021) or multiple object detection (Jia et al., 2020). Such shortcomings in threat modelling were observed and described early on (Sommer & Paxson, 2010; Gilmer et al., 2018), yet still only few work aim to understand AML in the real world. The first work in this direction, by Kumar et al. (2020), investigated which AML threats are feared in practice and reported that organization that apply AI are most concerned about poisoning. Other notable exceptions are the work by Boenisch et al. (2021) who, based on a questionnaire, compute an awareness score (based on ML, AML and other factors) for AML, which they find to be low. In contrast, Mirsky et al. (2021) reported that 24 of 33 offensive AI techniques pose a significant threat to organizations. They also observed that almost half of 102 queried cybersecurity organizations expect offensive AI techniques to manifest within the next 12 months.

Some AML incidents were also covered in the media.[1] Yet, this limits our understanding to incidents with public impact or incidents that happened to become public, a somewhat biased picture of all threats. Beyond this, Bieringer et al. (2021) found in their interviews first evidence for rudimentary AML attacks. We decided to address these potentially unreported cases and designed an anonymous questionnaire for ML-practitioners. In this short version of our work, we report the following insights from our 139 participants:

Contributions. (1.) We report evidence for occurrences of AML attacks, namely evasion and poisoning, in practice. (2.) Practitioners are concerned about AML, yet they also face privacy, general ML, and organizational challenges. (3.) Practitioners deem an AML attack as relevant for a complex array of reasons, including general business, financial, or even ethical concerns. When an attack is judged as irrelevant, our practitioners often reason that the application or deployment setting makes the attack infeasible.

Our results enable more detailed scientific investigations that encompass application and deployment when studying vulnerability. Furthermore, our insights are valuable when regulating and auditing ML systems, as we show that security (e.g., failures induced by an attacker) and benign failures are blurred, and as we analyze the underlying reasons for relevance or irrelevance of AML attacks.

2. Methodology

In this section, we describe the conception of our questionnaire. Afterwards, we discuss pretests, participant recruiting, the resulting sample and data preprocessing.

Questionnaire design. Our questionnaire contained open-ended questions, multiple choice questions, checkboxes, and relevance rankings based on Likert scales. For check-

[*]Equal contribution [1]University of Cagliari, Calgiari, Italy [2]QuantPi, Saarbücken, Germany [3]Eindhoven University of Technology, Eindhoven, The Netherlands [4]PluribusOne, Cagliari, Italy [5]CISPA Helmholtz Center for Information Security, Saarbücken, Germany. Correspondence to: Kathrin Grosse <kathrin.grosse@unica.it>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1] https://incidentdatabase.ai/?lang=en

boxes and multiple choice questions, we randomized the order to avoid order bias (Ferber, 1952). We first asked participants about their exposure AML to avoid priming when asking about the relevance of specific attacks later on. Afterwards, we inquired about the organizational and demographic background (see complete questionnaire in Appendix E). Questions, descriptions and wording of answer options were based on prior research.

AML background. AML encompasses a wide variation of threats that participants could be asked about in a questionnaire. We chose the six highest ranked attacks in the industrial ranking by Kumar et al. (2020). These threats are *poisoning*, where the attacker manipulates either samples (Rubinstein et al., 2009) or labels (Biggio et al., 2011) of the training data to reduce the accuracy achieved by the classifier at test time. In *evasion*, the attacker changes the test samples of a trained classifier (Dalvi et al., 2004; Szegedy et al., 2014) to trigger a wrong or specific output different from the original label. The attacker can also insert a specific pattern (trigger or *backdoor*) in the training set that has a strong association to one class (Chen et al., 2017). In *membership inference*, the attacker queries the model at test time to deduce whether a point was used in training (Shokri et al., 2017). The attacker can also query the model at test time, however this time to copy the model without consent (Tramèr et al., 2016), or *steal the model*. Kumar et al. (2020) distinguish model stealing and model extraction, a distinction that we avoided to obtain a simpler questionnaire. We also incorporate an impossible sanity-check attack to test that our participants are paying attention.

Pretests and recruiting. We implemented the questionnaire using Google Forms and ran a total of four rounds of pretests once there was the initial version of our questionnaire. The first three rounds with in total eight participants encompassed the full questionnaire. In the final round with three participants, we double-checked wording of some questions that were not sufficiently clear in the previous pretests. In the last round of feedback, no more necessary changes for the questionnaire emerged. Once pretests had been completed and the final questionnaire implemented, we started recruiting participants in the direct network of the first two authors of this paper. In doing so, we aimed to enable any necessary final adjustments to the questionnaire itself and to the way we approached participants before the study was widely advertised on social media channels. For detailed reasoning why the study is ethical, see Appendix A.

However, we found that direct messaging to both known and unknown possible participants came with higher conversion rates than general social media postings. Therefore, we joined several online communities for ML practitioners (e.g., R-Team for Data Analysis, Watson Developer Community, adversarial robustness toolbox, Data.Talks.Club) to approach potential participants via direct message on Slack. In doing so, we continuously monitored our sample with regard to representativeness to the overall target population. For example, the initial share of female participants in our study was below reported shares of female ML practitioners and we therefore explicitly targeted female communities.

Resulting sample. A total of 139 participants filled our questionnaire. More than two-thirds of the participants (71.2%) were male, 14.4% female, the remainder did not reply or did not want to disclose their gender. Albeit the sample is largely male, the percentage of female participants is comparable to reports studying the larger ML practitioner population (Kaggle, 2021). This accounts also for the distribution of participants' year of birth which was mostly between 1974 and 1996 with a median birth year of 1986. Also the distribution of academic degrees, with the largest group of master degrees (45.3%) mirrors this distribution.

With regards to participants' organizations, our sample is similar to industry surveys in ML with a significantly larger population (Kaggle, 2021). Albeit most of the participants' organizations were located in US and Europe (69%), our survey covers organizations from at least 26 countries.

Data pre-processing. Our questionnaire encompassed several possibilities for participants to reply with free text, for example in the question about feared threats or threat relevance questions. To analyse these replies, we applied four rounds of open coding. In each round, each coder assigned one or several codes to each participants statements. We then performed Strauss and Corbin's descriptive axial coding to group our data into categories and selective coding to relate these categories to our research questions (Strauss & Corbin, 1990). Throughout the coding process, we used analytic memos to keep track of thoughts about emerging themes. The final sets of codes are listed in Appendix F.

After coding, we computed annotator agreement. Given one document with many small text fragments, we opted for the Spearman correlation coefficient as a measure for annotator agreement (McDonald et al., 2019; Jinyuan et al., 2016) for the question about most concerning threats. This correlation coefficient, while not encompassing random overlap, allows us to take into account how often each code is used within the single document. For the relevance coding, we instead compute Cohen's kappa (Cohen, 1960), as we encode high and low relevance for each of the five attacks separately, yielding several documents with varying code assignment. We report the detailed agreement measures and code numbers in Appendix B. Given the semi-technical nature of our codebook, we consider our agreement substantial.

3. Empirical results

In this section, we analyze the responses of our participants. We first discuss which attacks occur in practice, which threats practitioners are concerned about, and finally investigate why attacks are deemed relevant or irrelevant.

3.1. Encountered AML threats

In contrast to (Mirsky et al., 2021), we find that only 17.2% of our participants estimate the likelihood of an AML attack within the next 12 months as high or very high. Instead, 49.6% as estimate the likelihood as low or very low. We thus aim to understand which threats were witnessed by our participants in practice. We had asked participants whether they had encountered a circumvention of their AI based workflows or systems. This was confirmed by 17.3% of our participants. More concretely, 5% (7 participants) witnessed one circumvention, 4.3% (6) two, 0.5% (1) three, 1.4% (2) four and 5.7% (8) more than four circumventions. To obtain more in depth knowledge, we asked participants to briefly describe the circumvention in a free text field. We now discuss the answers collected from our participants' replies.

AML in the wild. Of all the replies, 5 (3.6%) were AML threats. Three (2.1%) described an evasion attack. There were two kinds, the first one in relation to HR (1, "*users spam to optimize their strategy for job search*"), the second related to autonomous vehicles (2, "*autonomous vehicle image recognition errors leading to dangerous path planning*"). In the case of the latter two reports, participants doubted "*an 'intentional' circumvention*". Furthermore, there were two (1.4%) cases of poisoning. Whereas one remains vague, writing about "*ML systems being retrained to provide false outputs*", the second one very detailed, reporting that "*partner employees tasked with labeling training data feel threatened by automation, and either stall or sabotage the labeling effort, harming the models*".

Unclear replies. Further 9 replies, or 6.5%, contained no text, or replies like "*no details*" or "*brute force attacks*", that do not allow to deduce the exact circumvention. An additional six replies (4.3%) were data breaches. Whereas some referred on a high level to "*data privacy*", or "*incorrect data access*", others were slight more detailed: "*acquiring the Data for training AI Systems*". In these cases, we may assume, but cannot be sure that they are not AML related.

Circumventions not directly related to ML. Four (2.8%) descriptions were not ML related, but security threats, including resource theft (2, "*we got hit by crypto-miners pretty hard [...]*"), man in the middle attacks (1, "*a man in the middle attack between two workflows [...]*") and botnets (1, "*botnet communication*").

Additional mentions of circumventions. We later inquired about the relevance of specific AML attacks. In these replies, some participants reasoned that they had witnessed the threat already. One participant wrote for example, in the context of poisoning, "*however, something kind of like a poisoning attack happened, but was because of an over-prevalent family of malware that warped the model into performing worse than the last one. This did impact the deployment, but was because of a poorly configured filter not an attack.*" Another participant reported to "*[...] [have] evidenced during a penetration test scenario*" poisoning, evasion, and backdoor attacks. Another participant reports in the context of membership inference: "*we have seen users try to figure out what content will trigger our different abuse and spam identification models by trying different comment inputs and sharing these thoughts with others to help them bypass the potential identification.*".

Conclusion. There are occurrences of ML attacks in practice, namely poisoning and evasion. However, it is not always clear whether circumventions are malicious or benign performance failures. Furthermore, almost a third of our participants' replies remain vague, not allowing to reconstruct the exact attack. Finally, almost another third are data breaches, privacy, or other security issues.

3.2. Concerns about AML

The first question in our questionnaire aimed to understand what AML challenges practitioners face. To avoid priming, we had asked this question before we mentioned any other AML attack. Of all 139 participants, 93.5% provided a reply, and 22.9% provided more than one concern. In the following text, we refer to the number of codes assigned in agreement by both coders. As more than one code could be assigned to a reply, we report no percentages, as the total number of codes is not equal to the number of participants.

We tagged 21 times security challenges that were directly related to the AML, for example "*data poisoning*" or "*understanding the threats and associated risks of AI (and especially ML) - specific attack*". Several concrete AML attacks we later queried about, including poisoning (7), evasion (3), and model stealing (1) were named by our participants. However, most replies did not (only) contain AML threats. A few challenges, 10, were related to ML, for example "*explainable ML/NN*" or "*concept Drift*". There were also 16 challenges related to privacy. These encompassed "*data protection, Legal data collection, GDPR, Information security*", in other words both general privacy (10) concerns as also the challenges to be compliant with legislation (6).

We also found that 20 challenges concerned security in organizations. Corresponding replies are for example "*convincing stakeholders of the risks*", "*Protecting intellectual*

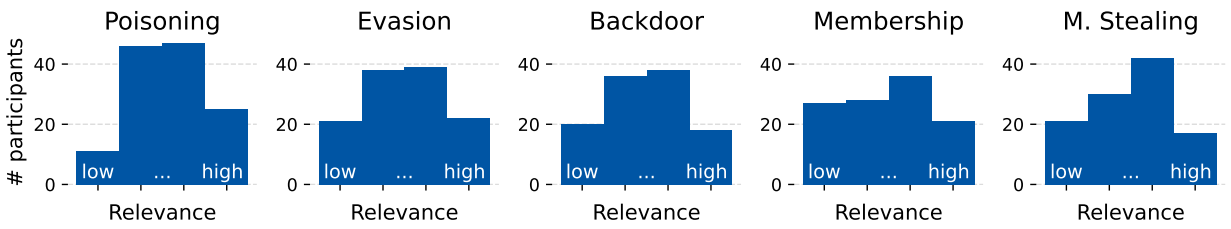

Figure 1. Reported relevance on a Likert-scale for the five AML attacks we presented to our participants

property" or "*achieving security guarantees while reducing false-positives*". They outline that challenges in AI can also encompass communication of risks (8), protecting intellectual property (7) or trade-offs that arise when both several factors are balances against each other (4). Furthermore, there were 35 challenges related to security unrelated to AML, including "*user access control*", or "*open Source Supply Chain (ie - NPM / Log4J vulnerabilities)*". One participant reasoned: "*hard to say but the traditional cybersecurity attacks are generally applicable in AI and those still seem to be most prevalent. [...] The adversarial scenarios as presented by evasion or poisoning are not as prevalent*", thus explaining why these replies are not about AML although we explicitly asked about it. The largest used group of assigned codes, 52, is related to data. While some of these replies are very vague (11, "*data leak*"; 17, "*data security*"), some are related to dealing with sensitive data (17, "*PHI/HIPPA*") or challenges when sharing data (8, "*the biggest difficulty is safely sharing data with others*"). In theory, almost all AML threats can be seen as attacks though data (through training for poisoning and backdoor, through test data for evasion, membership inference and model stealing). However, threats caused by data could encompass also non-AML threats including data quality, privacy, etc. We thus leave a detailed interpretation for future work.

Conclusion. There were few concrete mentions of AML threats. Although we explicitly asked for ML security, participants also raised other security and privacy concerns, reasoning that these were more pressing than AML. Concerns can be complex, and also encompass organizational challenges related to ML itself or risk communication or assessment. Finally, we find that often, participants reason vaguely about data security, leaving open whether this describes rather data quality, privacy, or AML issues.

3.3. Concern about AML threats

In this subsection, we analyze the arguments provided by our participants when reasoning that a threat is relevant or irrelevant. Previous work have studied factors on threat concern such as the ease to attack and defend, or possible benefit of carrying out the attack (Mirsky et al., 2021). We instead asked our participants without priming to give a short reason for the relevance or irrelevance of an AML threat. More concretely, we asked our participants how concerned they were about poisoning, evasion, backdooring, membership inference and model stealing and a sanity-check threat which is omitted here (see Appendix Cfor details). In this short version of the paper, we first discuss the high relevance of each poisoning, evasion, and model stealing, and then proceed in the same order with the arguments for the irrelevance of an attack. We summarize our results in Table 1, and plot the numerical relevance ratings in Figure 1. The discussion of the replies on backdoor and membership inference attacks can be found in Appendix D.

Poisoning–high relevance. The most frequent coded reply reasoning for relevance was the relevance within the applications setting of the participant (10 times, "*we use AI for security purposes, tampered training data is one of the best ways for attackers to evade the system*"). Following up codes are associated with relevance without argument (9, "*yes*"), and two codes associated with model performance (9 and 9). Participants also reasoned that an attacker was credible (5, "*sharing data across multiple users makes this a threat that needs to be considered*"), or that they understood the attack (7). Finally, some participants reported exposure to the attack (3), which is rarely the case for other attacks.

Furthermore, we found that 4 times, participants found the threat relevant as it would cause wrong decision making ("*models inform our decisions. Wrong models imply wrong decisions.*"). They furthermore reasoned that poisoning caused financial loss (3, "*altering training data could result [...] in catastrophic increased spending*") for their company or harmed fairness by potentially introducing bias (3).

Evasion–high relevance. The most frequent reply for high relevance of evasion was impact on model performance (11 times). At the same time, 6 participants reasoned that although evasion is relevant, it is not a security issue (6 times, "*it may be a case of overfitting*"). Further reasons included that evasion was easy to carry out (4), hard to defend (4), a threat relevant in the given application (3, "*attackers targeting our systems in this way may break them*"), or assumed to be relevant without providing an argument (4, "*it is*").

Table 1. Participants' reasoning for attack relevance. For each attack, we show the most frequent arguments (and their frequency).

Attack	Relevance	Irrelevance
Poisoning	Relevant in application setting (10)	(14) Data access control defense
	Impact on safety (9)	(9) Not relevant in application setting
	Impact on performance (9)	(8) Doubting attacker
	Relevant without further argument (9)	(5) Human in the loop defense
	Hard to defend (7)	(5) Unspecified defense implemented
Evasion	Impact on model performance (11)	(11) Data access control defense
	Impact on safety (6)	(10) Not relevant in use case
	Impact on decision making (4)	(7) Doubting attacker
	Easy to do (4)	(4) Defense in place
	Hard to defend (4)	(3) Hard to do
M. Stealing	Impact on intellectual property (8)	(13) Not relevant in use case
	General business impact (5)	(7) Querying access control defense
	Profit for competitor (5)	(4) Hard to do
	Financial loss (4)	(3) Doubting attacker
	Attacker credible (3)	(3) Model shortlived

As in poisoning, participants also reasoned that evasion affects decision making in their companies (4), or negatively affects fairness, bias, or ethics (3, "*brings in bias*").

Model stealing–high relevance. Most participants stated that model stealing results in a loss of their intellectual property (8, " *stealing IP*"). Further, participants reasoned that the attack was easy to do (4), was relevant in their application setting (3, "*it might lead to our models being reverse-engineered by clients.*") or the attacker had a motivation to carry out the attacks (3, "*when scraping enough data one could probably "copy" our models.*"). Practitioners also reasoned based on their understanding of the attack (4, "*technically its no brainer - It's very much possible.*").

Compared to other attacks, much more participants remark on the impact of model stealing. Several participants mention general business consequences (5, "*threat to the business*"), whereas others address profit for a competitor (5, "*would allow competitors to achieve our better results with minimum efforts.*"), financial loss (4, "*it costs a lot of money to train giant networks, hence the problem is very relevant in terms of investment*"), and business information leakage (3, "*could give unfair insights in our decision making*").

Poisoning–low relevance. The most frequent code (14 times) for irrelevance of poisoning attacks was that the data was not accessible to 3rd parties or the outside of the company ("*no one can access the training samples*"). Additional frequent codes are that the threat is not relevant in the considered use case (9 times, "*our training data comes [...] from clinical studies we conduct ourselves [...] so chances that someone interfere with the data gathering process are very low*") or doubting the attacker (8, "*we do not think any actor would be sufficiently motivated to attempt it*"). While some participants (3) also reason that the attack is simply hard to carry out, several state that a human in loop (5, "*the training data is curated by us*") prevents an attack or another defense they implement (5, "*very few publicly available data used for training*").

Evasion–low relevance. Most participants (11) arguing against evasion denied that an attacker could access the required test data. Almost as many reasoned based on their specific use cases (10, "*the podcast audio is stored with a number of distributors [...]. The corruption would have to occur amongst multiple distributors [...].*"). Many participants also doubted the attacker's motivation (7, "*[...] there would not be enough benefit to the actor*"). Further reasons included that the attack was hard to do (3), or that a defense was implemented (4, "*[...] the attack surface to alter data is minimized by multifactor access, role based access controls, time based tokens, logging, monitoring, and encryption.*"). Finally, we tagged some replies (5) as confused threat models because participants referenced training data ("*training data is usually high quality*").

Model stealing–low relevance. Most participants that do not see model stealing as a threat reasoning that it is not important in their use case (13, "*the use of the model requires domain knowledge so it's unlikely that someone outside the organization would be able to make a correct interpretation of it's functionality*"). Many participants further remark that their model are not accessible (7, "*we don't offer API's to our models.*"), or that their models are replaced within a short time-span and copying them yields no benefit (3, "*model is continuously updated, and previous models don't*

have much value"). Participants also reason that the attack is hard to carry out (4), generally irrelevant (3, "*this is a business model issue, not a technical issue*"). Participants also doubt that an attacker might benefit (3, "*the value of copying [our models] would be quite small for someone else*"), or reason that the attack does not apply in their deployment (3, "*the model is likely to be deployed on edge devices so it will be anyway known to the potential attacker.*").

Conclusion. The replies of our participants shed light on the complexity of AML in practical settings. While poisoning is the only attack where several participants report to have witnessed an attack, other attacks are deemed relevant due to their (potential) impact. Impacts are very diverse, and range from decreased model performance, wrong decision making, biased and unfair models, to business implications like leak of business information, financial loss and loss of intellectual property.

When an attack is deemed irrelevant, we find that often, the attacker would not have access to the data they require. In this sense, both application and deployment are orthogonal factors influencing vulnerability. The same use case may be security critical only depending on its deployment, and vice versa. Finally, we find that in some cases, the difference between a malicious decrease in accuracy and bad performance is not always well distinguished.

4. Limitations, future work, and implications

Before we conclude our work, we outline the limitations of our approach and discuss how to address them in future work and the implications of our findings.

Limitations. Although we attempted to make our study as anonymous as possible, many participants reporting circumventions of their AI did not provide details (Section 3.1); one potential participant candidate even denied participation upfront for confidentiality reasons. This might imply a potential bias in our study, leading to our results underestimating the real occurrence of circumventions in practice. In addition, these concerns raise the question how ML security in practice can be studied in detail beyond studies as ours and the threat incident database[2]. One possibility for future work might be to simply ask for the circumvention, and no other information, for example. This limits the usability of the results, but does allow conclusions as long as participants provide sufficiently many details.

Future work. It would thus be beneficial to conduct a larger study which is purely based on threat exposure. Furthermore, and orthogonally, we find that not only the application, but also deployment is a crucial factor dictating vulnerability of an ML system (Section 3.3). Both factors need to be monitored and can then jointly with for example exposure and AI maturity be used to assess risks in practice. Such an assessment is also helpful to understand how high the risk of an AML attack is truly—as our 16% exposure does not take into account cases where an attack would be virtually impossible due to implemented access control, for example.

Implications. Such insights are also crucial when regulating, auditing or carrying out threat modelling for ML in practice. We found that when discussing both pressing security issues (Section 3.2) and the relevance of attacks (Section 3.3), participants sometimes attributed attack consequences to benign failure cases of ML. There seems to be no consensus about the boundary between attacks (e.g., malicious degradation of performance) and benign degradation of performance. Current existing standards, like for example ISO26262 about vehicles, originate in deterministic systems and software development. This absence of a treatment for benign failure cases has been identified as one of many shortcoming when applying them to ML-based autonomous cars (Salay et al., 2018). Yet, it remains an open question whether benign failure cases and AML attacks *need* to be distinguished in practice. Future work need to determine whether such a distinction is relevant or necessary.

Furthermore, we provide insights into why practitioners think specific attacks are relevant or irrelevant (Section 3.3). These insights could be a starting point for educational measures by regulators, auditors or in general organizations that deploy AI to generate awareness for possible attacks. More concretely, our results could help educating business stakeholders that they, for example, have to consider model stealing because it is a potential target for IP theft, or that they should consider poisoning as it may affect their decision making. Finally, our findings also help to understand which factors have to be taken into account when threat modelling an AI application.

5. Conclusion

We found evidence for AML attacks, more specifically evasion and poisoning, in practice. However, it remains often unclear whether circumventions are malicious or benign. In addition, also privacy, ML, and organizational challenges like data drift are of importance to our participants. We furthermore find that the presence or absence of concern for an AML attack is complex, encompassing factors such as financial loss, ethical concerns, decision making, but also application setting and the way in which ML is deployed. Our results yield important insights for regulators and auditors as we analyze relevance and irrelevance, and point out that the boundary between malicious and benign failure cases is not always clear. We are further confident that we are contributing towards more research that elicits when ML systems are vulnerable in practice.

[2] https://incidentdatabase.ai/?lang=en

Acknowledgements

The authors are deeply grateful to all our pre-testers and participants. We would further like to thank Beat Busser, Federico Marengo, Brian Pendleton, and Jessica Rose for helping with the recruitment. This work was supported by the Province of Upper Austria within the COMET program managed by FFG in the COMET S3AI module.

References

Barreno, M., Nelson, B., Sears, R., Joseph, A. D., and Tygar, J. D. Can machine learning be secure? In *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, pp. 16–25, 2006.

Bieringer, L., Grosse, K., Backes, M., and Krombholz, K. Mental models of adversarial machine learning. *arXiv preprint arXiv:2105.03726*, 2021.

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

Biggio, B., Nelson, B., and Laskov, P. Support vector machines under adversarial label noise. In *ACML*, pp. 97–112, 2011.

Boenisch, F., Battis, V., Buchmann, N., and Poikela, M. "i never thought about securing my machine learning systems": A study of security and privacy awareness of machine learning practitioners. In *Mensch und Computer 2021*, pp. 520–546. 2021.

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

Cinà, A. E., Grosse, K., Demontis, A., Vascon, S., Zellinger, W., Moser, B. A., Oprea, A., Biggio, B., Pelillo, M., and Roli, F. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *arXiv preprint arXiv:2205.01992*, 2022.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

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

Ferber, R. Order bias in a mail survey. *Journal of Marketing*, 17(2):171–178, 1952.

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

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

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

Jia, Y. J., Lu, Y., Shen, J., Chen, Q. A., Chen, H., Zhong, Z., and Wei, T. W. Fooling detection alone is not enough: Adversarial attack against multiple object tracking. In *International Conference on Learning Representations (ICLR'20)*, 2020.

Jinyuan, L., Wan, T., Guanqin, C., Yin, L., Changyong, F., et al. Correlation and agreement: overview and clarification of competing concepts and measures. *Shanghai archives of psychiatry*, 28(2):115, 2016.

Kaggle. State of machine learning and data science, 2021. URL https://storage.googleapis.com/kaggle-media/surveys/Kaggle's%20State%20of%20Machine%20Learning%20and%20Data%20Science%202021.pdf.

Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., and Xia, S. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 69–75. IEEE, 2020.

McDonald, N., Schoenebeck, S., and Forte, A. Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW): 1–23, 2019.

Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Lee, W., Elovici, Y., and Biggio, B. The threat of offensive ai to organizations. *arXiv preprint arXiv:2106.15764*, 2021.

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

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

Rubinstein, B. I., Nelson, B., Huang, L., Joseph, A. D., Lau, S.-h., Rao, S., Taft, N., and Tygar, J. D. Antidote: understanding and defending against poisoning of anomaly detectors. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pp. 1–14, 2009.

Salay, R., Queiroz, R., and Czarnecki, K. An analysis of iso 26262: Machine learning and safety in automotive software. Technical report, SAE Technical Paper, 2018.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *IEEE S&P*, pp. 3–18, 2017.

Sommer, R. and Paxson, V. Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE symposium on security and privacy*, pp. 305–316. IEEE, 2010.

Strauss, A. and Corbin, J. *Basics of qualitative research*. Sage publications, 1990.

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

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. Stealing machine learning models via prediction apis. In *USENIX*, pp. 601–618, 2016.

Woitschek, F. and Schneider, G. Physical adversarial attacks on deep neural networks for traffic sign recognition: A feasibility study. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pp. 481–487. IEEE, 2021.

A. Detailed ethical reasoning for our study

Our study encompasses human participants, which always raises the question whether the study is ethical. In our case, participation is purely voluntary, there are no financial incentives as we did not pay our participants. We only require anonymous data, and our forms always gives participants the option to not reply to a question within the questionnaire. We further do not deceive or participants as part of the study design. The risks imposed on our participants are thus small. On the other hand, we gain important insights for the field of the adversarial machine learning community. We are thus confident that our study design is indeed ethical.

B. Interrator Agreement

We depict the detailed interrator agreements in Table 2.

C. Sanity check threat

We also asked about one additional sanity-check threat ("*Altering training data to delete an untrained model. In other words, the training data contains a pattern that will delete the model after training.* "). Although some participants reported high concern about this threat, a Mann-Whitney-U test confirmed statistically significantly lower ratings compered to all other threats ($[1.4e^{-10} < p < 1.2e^{-16}]$). We thus do not discuss this threat in the discussion and evaluation.

D. Omitted results about backdoors and membership inference

We now discuss the two attacks that we omitted in the main paper. As in the main text, we first review the high relevance replies for backdoors and membership inference and then discuss why participants argued that attacks are not relevant.

Backdoors–high relevance The most frequent (9) codes in Backdoor relevance were associated with model performance. However, in many cases (6), participants argued that the attack was not a security, but a benign performance issue ("*because this may happen unintentionally, and you never know what happens*"). Often, Participants reasoned that backdoors induced bias (6, "*that's intentional deception inducing malicious bias, altering fairness of the model*"). Several participants brought forward that the understood the underlying mechanism (6, "*it is like adding some conditions on prediction*"), had read about the attack (3), reasoned it was hard to defend (3), or gave no further argument (3).

Intriguingly, only in total four replies were tagged consistently by us as impact-based arguments for relevance. This stands in sharp contrast to all other attacks, where more such arguments were made.

Membership Inference–high relevance Most participants argued that they were concerned about the resulting data breach (21, "*the possibility of de-anonymizing data would be a concern that can't be understated*"). Some participants understood the underlying mechanism (4, "*it allows someone to reverse engineer the inputs and potentially identify where the data came from as well as who or what is/isn't included*"), other reasoned that the threat was relevant in their specific use-case (3, "*especially our model could be queried to generate training data*") or did not give additional arguments (3).

Our participants also reasoned that Membership inference causes business information leakage (3, "*could be relevant because*

Table 2. Inter-coder agreement of free text replies. We compute Spearman correlation (*) and Cohens Kappa (†) for the most feared threat (first question) and the replies of high/low relevance of the five investigated threats. The amount of total codes is the maximum number of codes given from one of the two coders. The discussion of Backdoor and Membership attacks can be found in Appendix D.

		Poisoning		Evasion		Backdoor		Membership		M. Stealing	
AML concerns		high	low	high	low	high	low	high	low	high	low
Agreement	.96*	.79†	.65†	.77†	.69†	.74†	.55†	.67†	.53†	.62†	.55†
Total codes	232	86	61	67	63	47	57	54	47	53	47
# disagreeing	37	6	8	7	8	7	10	8	11	10	9
# uncoded	-	2	2	3	-	4	4	2	4	4	1
# replies	136	69	51	54	49	43	49	45	45	45	40

it would allow our clients to get information about the competition they would normally not have.") or noncompliance with existing regulations (3, "*GDPR requires that I don't accidentally leak data that was supposed to remain private*").

Backdoors–low relevance Most participants, to argue that backdoors were irrelevant, referred to a non-credible attacker (14, "*I can't see how it would benefit an attacker or user to be able to get a certain prediction from our products*"). Almost as many participants (12) reasoned that an attacker would not have access to the required data (12, "*we do not expose our training data outside our company*"). Furthermore, several (5) participants reasoned that backdoors were not relevant for their use case ("*specificity of the application*"). Some (3) participants reasoned another threat was more likely ("*higher cost to an attacker than efficient traditional attacks*") or they had a defense in place (3, "*logging and audit trails for user actions would help identify any customer bad actors*").

Membership inference–low relevance To reason for the irrelevance of membership inference, participants often referred to their specific use case (10, "*we work on new data in news and the probability of that happening since our models are trained in old data is very unlikely*") or directly stated they were dealing with non-sensitive data (9, "*the training data is publicly available anyway*"). In addition, participants sometimes did not provide an additional argument (4), doubted the attacker (3, "*for our use cases, I can't (yet) see how anyone would stand to gain from this*") or reasoned their model was not accessible at test time (3, "*the model cannot be queried directly by the users*").

E. Questionnaire

Survey on security of applied AI Thank you for taking the time to give us your perspective on the practical state of security in the context of applied AI. We are looking for all kinds of circumventions in the context of AI, like altered data to force wrong classification outputs, manipulated data to change trained models, or simply user collected input data that is in any way harmful to an AI model. This research project is a collaboration between **blinded for submission** Your participation will inform the design of a framework to adequately approach security of applied AI. Completing this survey will take about ten minutes.

This study consists of three parts. Part I addresses security within your AI workflows, products or systems. Part II aims at exploring AI practices in your organization. Part III is about your individual background.

We value your privacy! This study is anonymous and we do not collect any information that can be traced back to you or your organization. Aggregated and anonymized results are processed according to the European General Data Protection Regulation (GDPR) and might be published in a scientific venue. By clicking 'Next', you agree to this use of the information you provide. The research team is at your disposal for any privacy-related **Questions via *author's email-address***.

Click 'Next' to start with Part I of the study.

Part I - Security of AI within your organization

Question 1: In your daily work and your organization's AI workflows, products or systems - what are the most pressing security challenges?
(text field to type reply)
Question 2: Did you already experience a circumvention of your AI based workflows, products or systems?
yes/no
IF YES:
Question 2.1: How many circumventions of your AI based workflows, products or systems have you experienced?
1,2,3,4,>4
Question 2.2: Please describe the most severe circumvention of your AI based workflows, products or systems.
(text field to type reply)
Question 3: 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?
(linear scale from 1 (very low) to 5 (very high))
Question 4: Which of the following approaches does your organization implement in terms of the security of your AI based workflows, products, or systems?

None

Documentation: Usually printed instructions, comments, and information for using a particular AI system or AI specific hardware.

Guidelines: Codes of conduct or policies for AI security.

Mitigations: Implemented processes to make threats for AI systems less severe, dangerous, painful, harsh, or damaging.

Fail safe plans: incorporating some feature for automatically counteracting the effect of an anticipated possible source of failure within or around an AI system.

Human in the loop: Natural person that oversees an AI system.

Incident response: Measures that determine when an attack on an AI system has occurred or is underway and what should be done about it.

Security testing: Resilience testing of AI systems with regards to unauthorised third parties altering their use, performance or exploiting system vulnerabilities.

Other: (textfield)

You will now be confronted with descriptions of specific threats to the security of AI. Please think about how these threats might take effect in your AI workflows, products, or systems.
Question 5: Do you consider the following threat scenario relevant in your work?
(placeholder for attacks, see below)
very relevant; relevant; not very relevant; irrelevant; I don't know; I don't understand threat scenario
Question 6: Why do you think this threat scenario is (placeholder for previous selection)?
(text field to type reply)
These 2 questions are repeated iteratively for the attacks we want to test:

1. **Q7,8**: Altering training data to harm model performance during deployment. In other words, the model is optimized on tampered training data, which affects the resulting model.
2. **Q9,10**: Altering test data to harm model performance during deployment. In other words, the trained model is presented with specially crafted inputs that lead to wrong predictions.
3. **11,12**: Altering training data so that the model outputs a chosen class whenever a particular pattern is present in the input data. In other words, altering the training data to contain a certain association between a pattern and a label, the resulting model contains a backdoor.
4. **13,14-Sanity**: Altering training data to delete an untrained model. In other words, the training data contains a pattern that will delete the model after training.
5. **15,16**: Given input data and the predictions of a model, determine whether the given data sample is part of the training data. In other words, the model is queried to obtain crucial information about the used training data.
6. **17,18**: Given an API / black box access to a model, copy its functionality. In other words, repeatedly observe input and output pairs from the model to reproduce its functionality.

Part II - AI within your organization

Question 17: In which country is your organization headquartered?
(drop down with all 195 countries of the world)
Question 18: What is the number of employees at your organization?
$<$10, 10-49, 50–99, 100–249, 250–499, $>$500
Question 19: Which industry area describes your organizations best?
Customer Service & Support, IT Security, Production, Marketing, Computer Audition, Research, Forecasting, Computer Linguistics, Computer Vision, Agriculture Forestry & Fishing, Finance & Insurance, Arts Entertainment & Recreation, Manufacturing, Water & Waste, Healthcare, Retail & Commerce, Transportation & Mobility, Other
Question 20: What kind of data analysis do you work with primarily?

Supervised Learning - input data is presented alongside with labels for this data

Unsupervised Learning - only input data is given, without labels

Semi-supervised Learning - the data is partially labelled

Reinforcement Learning - task of finding suitable actions in a given situation in order to maximise a reward

Other: (textfield)

Question 21: What do you use AI for primarily (e.g. sentiment analysis, object detection, malware classification)? (text field to reply)
Question 22: What input data do you work with primarily? (tick most specific)
Images, Videos, Speech/Audio, Text/Documents, Network traffic, Social media data, Files/Source Code, Other: (textfield)
Question 23: What kind of labels do you work with primarily?

Unlabelled data

Categorical (for example 'cat' or 'noun')

Real values (for example '1.4' or '1.8')

Structured data (for example bounding boxes)

Other: (textfield)

Question 24: 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

Models in production, for 2-4 years

Models in production, for >5 years

Question 25: Which of these goals are part of your organization's ML-model checklist?

Performance: Degree of accuracy and/or conformity of the AI to the ground truth or provided labels

Fairness: AI's lack of favoritism toward one feature value or another

Explainability: The ability to explain the rationale behind AI's decision

Security: measures taken to guard against espionage, sabotage, crime, or attack of the AI

Privacy: freedom from unauthorized intrusion of the AI or its data

Ethics: moral issues or aspects (such as rightness) that concern the AI

System Response time: time delay between a user's initiation of a command on an AI and the system's task completion

Other: (textfield)

Part III - Demographics and your AI background

Question 26: In which year were you born?
(2021-1935)
Question 27: What gender do you identify with?
Female, male, other, I do not want to disclose
Question 28: In which country are you located? (drop down with all 195 countries of the world) **Question 29**: What is your level of education? Please specify the highest.
Highschool, Bachelor, Master / Diploma, Training / Apprenticeship, PhD, Other: (textfield)
Question 30: What is your role in your team?

ML Engineer: You build, deploy or improve ML models.

ML Researchers: you develop new algorithms

Data Scientist: You deploy ML models to extract insights from data.

Domain Expert: You rely on ML to improve or accelerate decision-making in a specific application domain.

Product Owner: You manage the incorporation of user needs into your organization's ML-based products.

Auditor: You review ML systems to ensure their compliance with regulatory requirements.

Other: (textfield)

Question 31: Please complete the following sentence. When it comes to machine learning, I believe I have...

No knowledge: I might be aware of ML, but have no knowledge about it

A little knowledge: I know basic concepts in ML, but have never applied it

Some knowledge: I have applied ML concepts at least once before

Moderate knowledge: I apply ML concepts somewhat frequently for my work, class, or leisure

High knowledge: I apply ML concepts very frequently or create cutting edge Software

Question 32: In which of these areas have you taken a lecture or intense course?
None, Machine learning, Security, Adversarial Machine Learning

Submission
Thank you for giving us your valuable perspective on AI security. In case you want to reach out to the research team, give feedback or receive the results of this study, please feel free to do so via *authors mail addresses*

F. Complete sets of Codes

We here depict the full sets of codes used for free text replies. This encompasses the question about the most pressing AML security challenge (Table 3) and the relvance reply encoding (Table 1).

Table 3. Codes used to encode the first question, where participants describe their current AI security concerns.

Group	Code	Group	Code	Group	Code	Group	Code
AML	General	Non AML	General	Privacy	General	ML	General
	Poisoning		Libraries		Regulations		Explainability
	Evasion		Access				Bias
	Model stealing		Customer	Data	General		Concept drift
	Performance impact		Code breach		Data sharing		
	Robustness		3rd party provider		Breach	Organization	Complexity
	Test time		Precise threat		Sensitive data		IP
	Training time		Cloud		Classify if sensitive		Trade offs
	Model itself						Security awareness
							Human harm

Table 4. Codes used to encode the relevance replies, where participants argue why (or why not) they think an AML attack is relevant or not.

Group	Code	Group	Code	Group	Code
Relevance	General relevance	Impact	General business	Defense	Easy to defend
	General irrelevance		Financial loss		Hard to defend
	Easy to do		Business information leakage		Data access control
	Hard to do		Profit for competitor		Model acccess control
	Has encountered threat		Intellectual property		No sensitive data
	Has not encountered threat		Reputational damage		Model shortlived
	Attacker credible		Regulatory compliance		Human in the loop
	Doubting attacker		Data breach		Implemented
	Relevant in application setting		Wrong decision making		
	Not relevant in use case		Human harm		
	Not relevant for deployment		Ethics/fairness/bias	Perception	Did not understand threat scenario
	Understands attack mechanism				Confusion across threat models
	Theoretical exposure to threat				
	Other threat more likely				
	Safety				