Machine Learning Security in Industry: A Quantitative Survey

Kathrin Grosse, Lukas Bieringer, Tarek R. Besold, Battista Biggio, *Senior Member, IEEE,* Katharina Krombholz

***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 report on a quantitative study with 139 industrial practitioners. We analyze attack occurrence and concern and evaluate statistical hypotheses on factors influencing threat perception and exposure. Our results shed light on real-world attacks on deployed machine learning. On the organizational level, while we find no predictors for threat exposure in our sample, the amount of implement defenses depends on exposure to threats or expected likelihood to become a target. We also provide a detailed analysis of practitioners' replies on the relevance of individual machine learning attacks, unveiling complex concerns like unreliable decision making, business information leakage, and bias introduction into models. Finally, we find that on the individual level, prior knowledge about machine learning security influences threat perception. Our work paves the way for more research about adversarial machine learning in practice, but yields also insights for regulation and auditing.**

***Index Terms*—Adversarial Machine Learning, Machine Learning Security, Quantitative User Study.**

I. Introduction

A large body of academic work focuses on machine learning security or adversarial machine learning (AML) [1]–[11]. These works investigate how machine learning (ML) can be circumvented and exploited by an attacker. For example, an attacker can tamper with the training data, yielding a model inferior in performance or that is sensitive to attacker specified, small parts of the input [4]. Alternatively, the attacker slightly alters test data to change the output of an ML model [5], [10]. In addition, an ML model may leak the used training data [12] or can easily be copied when freely exposed [11].

Many of the settings studied in ML security can be criticized for being rather artificial. However, already these settings are hard to solve [4], [13]. One possible cause is that even though the current usage of ML in security and threat modelling have been criticised [14], [15], there is little work on ML security in the real world. In the first work in this direction, by Kumar et al. [16] investigated which AML threats are feared in practice by interviewing 28 organizations whose largest concern were poisoning attacks. Mirsky et al. [17] reported that the 22 interviewed organizations perceived 24 of 33 offensive AI techniques as a significant threat. Moreover, Bieringer et al. [18] found evidence for rudimentary attacks on AI in the wild in their 15 qualitative interviews with ML practitioners. Boenisch et al. [19] found that in their 83 participants, security and ML security awareness of ML practitioners was overall low. In contrast to these previous works, our sample with 139 participants is larger and more diverse. Our focus only on ML security (not considering privacy, or offensive AI) allows us further to study threat concern in depth, and to run statistical tests on our participant's replies. Finally, we are the first to publish an estimate about ML security incidents from the real world but not covered in media.

More specifically, to shed light on the state of AML in practice and the factors influencing organizations' approach to ML security, we conduct a quantitative survey among ML practitioners. Inspired by prior work [17], we investigate which threats are dreaded and why. Furthermore, we control for variables like application area [19], how long ML has been used in production, data type, and prior knowledge [18]. All these questions and variables form part of our anonymous questionnaire for ML-practitioners. Our 139 participants help us to shed light on the following topics:

AML in practice. We find that there are occurrences of AML attacks, more specifically evasion and poisoning, in practice. However, non-ML security threats (e.g., access control, botnets, resource theft, etc.) are also prevalent and (still) seem to pose a larger concern, together with organizational challenges, privacy and benign ML challenges.

AML within organizations. We find that exposure to AML threats is not related to organization size, time that ML is used in production, or organization area. Yet, some of these factors, together with exposure, influence strongly how many mitigations an organization has in place. Furthermore, we investigate why practitioners deem an AML attack as relevant, and find a complex array of reasons, including business or financial, even ethical concerns. When an attack is judged as irrelevant, this is often a consequence of an application or deployment setting that makes the attack infeasible.

AML for practitioners. We find that self reported prior knowledge, in particular in AML, increases the concern reported for individual attacks. Concerning gender, we find that two of the five attacks tested are rated significantly less important by women, the remaining threats are rated similarly.

Our results open the avenue of more in depth studies that encompass application and deployment when studying vulnerability. Our results also shed light on the relationship between knowledge and threat perception. Our insights are furthermore valuable when regulating and auditing ML systems, as we a) analyze the underlying reasons for relevance or irrelevance of specific AML attacks and b) show that security and safety are

First two authors contributed equally. K. Grosse is with the EPFL; L. Bieringer with QuantPi; T.R. Besold with Eindhoven University of Technology; K. Krombholz with CISPA Helmholtz Center of Information Security; and B. Biggio with the University of Cagliari.

conflated by our participants. The latter refer to the difference of benign system failures (safety) and attacker induced failures (security), as used in system analyses [20]. We finally deduce that ML security incidents in practice are not as common as for example non-AML security, but that monitoring ML security might be beneficial.

II. Background

Before we describe our questionnaire and the methodology of our study, we would like to provide background on ML security or AML, as we confronted participants with the most relevant attacks discussed in AML theory [2] and practice [16]. More concretely, we focus on the six most relevant attacks in the industrial ranking by Kumar et al. [16]. We now give a rough overview of these attacks, and refer the interested reader to Bieringer et al. [18] or recent surveys [2], [4]. To ease understanding of the below attacks, we first define ML formally. In ML, a by ω parametrized function f is optimized during training to fit the training data X, Y. After training, $f(X, \omega) = Y$, and we expect the classifier to generalize to unseen test data, e.g. $f(X_t, \omega) \approx Y_t$. We further write X^* when an attacker has altered, or perturbed, the benign data X.

Poisoning. Poisoning affects, via the training data, the classifier at test time to reduce the overall performance or accuracy. To this end, the attacker manipulates samples X^* [21], labels Y^* [22] or both before training. Poisoning defenses are, compared to most attacks, well understood in terms of trade-offs, for example when it comes to attack strength and detectability [4]. We also investigate backdoors, a variant of poisoning which affects, via the training data $\{X^*, Y^*\}$, a specific subset of the test data $\{X_t, Y_t\}$ [3].

Evasion. Evasion or adversarial example affect, via the perturbed test data X_t^*, the classifier f and forces it to output either a predefined or a wrong output for an input, hence $f(X_t^*, \omega) \neq Y_t$. To this end, the attacker changes the test input of a trained classifier carefully based on the classifier [5], [10]. Most evasion defenses are caught in an arms-race [13].

Membership inference. While the previous attacks harmed performance, this attack harms the privacy of the training data X. More specifically, the attacker queries the model at test time to deduce whether a point was used in training, e.g. $x \in X$ [12]. Against these attacks, defenses have been proposed [23], but they are not as well understood as poisoning or evasion, for example.

Model stealing. Finally, in model stealing, the attacker harms the intellectual property of the model owner by copying the model f without consent. To perform this attack, the attacker queries the model with the goal to use the obtained data to steal the model [11]. For this attack as well, defenses have been introduced [24], nut no consensus has been reached so far on a standard defense.

Kumar et al. [16] distinguish model stealing and model extraction, a distinction that we avoided to decrease the risk of confusion for our participants. Beyond evidence by Bieringer et al. [18], Lin and Biggio [25] and Kumar et al. [16], few works have studied the relevance of these attacks in practice.

III. Methodology

Given this limited knowledge about AML in practice, we opted to design an questionnaire study that allows to reach many participants, yet still allows to process open text answers concerning for example the relevance of attacks. In this section, we describe our initial estimates on participants and how we designed our questionnaire.

Power analysis. Before we started designing our questionnaire, we needed to know how many participants we required: few participants would allow a longer questionnaire, more required a shorter. Many of our research hypotheses were testable with ordinal regression, as we planed to investigate how ordinal factors like organization size or organization maturity influence exposure to threats, for example. To use this regression with a power of 0.8, a medium effect and a significance level of 0.05, we would need, depending on the number of predictors, between 67 (2 predictors) and 97 (6 predictors) participants. Green's rule of thumb provides similar results. An alternative for two nominal variables is the Mann-Whitney-U test. This test assume as H_0-hypothesis that the two samples follow the same distribution, and is sufficiently powerful already for small sample sizes of 20 [26].

A. Questionnaire design

Given the required sample size of roughly 100, we opted for an anonymous survey with in total 32 questions. The questionnaire contained open-ended questions, multiple choice questions, checkboxes, and relevance rankings based on a Likert scale. For checkboxes and multiple choice questions, the order of all replies was randomized to avoid order bias [27]. Questions, descriptions as well as the wording of answer options for multiple choice questions were based on prior research. In the following, we detail references used for the questionnaire along it's three parts, (1) AML in practice, (2) organizational background of participants, and (3) individual background of participants. The complete questionnaire can be found in App. A.

We decided not to pay our participants to avoid money-driven participation. Furthermore, To not restrict our participants to ML specifically, we used the term artificial intelligence (AI) throughout the questionnaire, although our analysis focuses on ML security.

Questionnaire–AML in practice. The first part of our questionnaire addressed security within participants' AI workflows, products or systems. This included an open-ended question about the most pressing security challenges in participants' daily work and an indication on whether they had already experienced a circumvention of AI based workflows, products or systems [18]. In addition, we asked participants to estimate their risk to become a victim of attacks on their AI based systems during the next 12 months and to provide information about their organization's approaches towards AI security [28]. Choice and concrete wording of measures for AI security have been based on prior research [29], dictionaries,[1] recent

[1] https://www.merriam-webster.com

regulatory approaches,[2] and auditing frameworks.[3] Following Huaman et al. [28], we further inquired the relevance of the previously discussed relevant ML attacks [2], [16]. For each attack and a made-up sanity threat, participants were shown a description (see App. A) to assess the attack's relevance.

Questionnaire–organizational background. We also queried information on participants' organizations and their AI practices. This included basic information like the number of employees [30] or industry area [16]. With regards to AI practices, we queried information with regards to organizations' primary data analysis type (supervised/unsupervised learning, reinforcement learning) and input data (images, program code, etc.) and labels (real valued, discrete, etc). In addition, we asked for the status of AI projects in the organization, for example "*evaluating use cases*" or how many years the organization had already models in production. This is commonly referred to as an organization's AI maturity[4]. We also asked participants for goals within their organizations ML-model checklist [31], [32].

Questionnaire–demographics. The last part of the survey was about the individual background of our participants. It covered basic demographic questions, but also a description of participants' roles within their teams to address practitioners with formal and instrumental knowledge [33], and to include a broader AI audience [34]. To test for a possible relevance of field related expertise, we also queried participants' self-reported knowledge in ML and asked whether they had taken any lecture in ML, security or AML [18].

B. 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 this 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.

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, and we therefore explicitly targeted female communities. Our initial power analysis required more than 97 participants. Taking into account that not all participants reply to all questions, and having reached 104 participants after two and a half months, the authors decided to recruit >125, yielding the eventual sample size of 139.

[2]e.g. Artificial Intelligence Act by European Commission

[3]e.g. AI Cloud Service Compliance Criteria Catalogue by German Federal Office for Information Security, AI Auditing Catalogue by Fraunhofer IAIS

[4]https://info.algorithmia.com/hubfs/2019/Whitepapers/The-State-of-Enterprise-ML-2020/Algorithmia_2020_State_of_Enterprise_ML.pdf, page 8.

IV. Sample and Data Pre-Processing

We now discuss our sample encompassing the 139 participants and the pre-processing of the free-text replies. We first review the individual background of our participants, and discuss gender, age, education and the professional role. We then focus on the organizational background, and discuss organization areas, organization location, size, and concrete AI usage. Afterwards, we discuss the detailed procedure how we analysed and encoded the free-text replies and the agreement that we obtained across the two coders.

A. Sample description

A total of 139 participants filled our questionnaire, with additional 5 participants submitting empty forms (total 144). In addition, nine participants whom we contacted reported to have had a look at the questionnaire but did not want to participate as they felt they did not have enough knowledge, had not been exposed to the topic or felt the topic was not relevant in their area. One additional participant denied to take part due to confidentiality reasons.

Individual background of participants. Of our 139 participants, more than two-thirds (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 [35]. The distribution of participants' year of birth was mostly between 1974 and 1996 (median 1986, see Fig. 1), and is also similar [35]. Also the distribution of academic degrees, with the largest group of master degrees (45.3%) roughly mirrors this distribution [35]. Beyond general education, only few participants self-reported none or little knowledge in ML (5.7%). Many reported moderate (39.5%) or high knowledge (40%). More specifically, in a question asking for ML, AML and security knowledge, over three quarters of our participants reported ML knowledge (82%), only a third reported to be knowledgeable in security (34.5%) and even less in AML (28.7%). Less than a fifth reported knowledge in all three areas (17.3%). The most frequent combination was ML and security (30.2%), then AML and ML (28%), then AML and security (18%). Finally, the most frequent role in team is ML engineer, which almost a fourth (23.7%) indicated. The two second largest groups, both roughly one fifth, were ML researchers (20.9%) and data scientists (20.1%). Our sample also encompasses a few rather less technical roles, including nine domain experts (6.5%), five auditors (3.5%), and three product owners (2%). Roughly a fifth of the participants (18%) preferred to specify

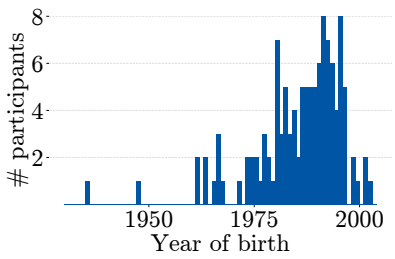


Fig. 1: The age distribution (years of birth) of our participants.

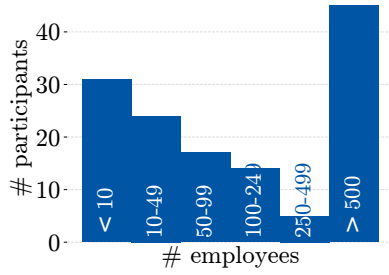


Fig. 2: Grouping our participants' organizations according to size.

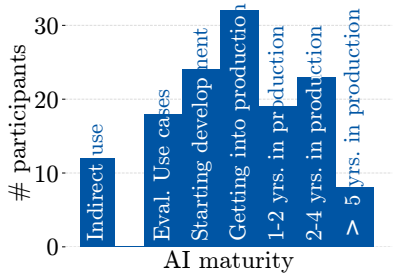


Fig. 3: Self-reported AI maturity of participants' organizations.

their own role, and named for example roles like "*consumer*", "*technologist*", "*consultant*", or "*CEO*".

Organizational background of participants. These roles were filled in organizations of overall diverse areas, with the largest two groups being healthcare (12.9%) and IT security (10.8%). Further areas included, but were not limited to marketing (6.5%), computer vision (5.7%), and finance and insurance (5%). Albeit most of the participants' organizations were located in US and Europe (69%), our survey covered organizations from at least 26 countries (roughly nine out of ten participants (87%) provided a country). More than a fifth of organizations participating in this study had 1-49 employees (22%) while 45 (32.3%) participants were working for organizations with more than 500 employees (see also Fig. 2), similar to existing industry surveys in ML [35]. With regards to the AI maturity of their organizations, 24 (17.2%) participants stated that their organization was starting to develop models whereas 32 (23%) reported that their organization was getting developed models into production. Albeit 42 (36.6%) participants reported that their organization had models in production for 1-4 years, relatively few subjects (5.7%) reported that their organization had ML models in production for more than 5 years (see Figure 3). Concerning the concrete usage of ML techniques, more than half of our participants stated they used supervised learning (56.8%). Significantly less, about a fifth, used semi-supervised learning (19.4%), less unsupervised learning (14.4%) and few reinforcement learning (6.5%). Even fewer reported to work with all four categories of learning (3%). Related to the previous question, two thirds of our participants used categorical labels (67.6), roughly half structured labels like bounding boxes (51%) or real valued labels (49%). Less worked with unlabelled data (41.7%), and almost a fifth reported to work with all kinds of labels (18%).

B. Data pre-processing

Our questionnaire encompassed several possibilities for participants to reply with free text, for example in the question about dreaded threats or threat relevance questions. To analyse these replies, the first two authors of this paper applied four rounds of open coding. In each round, each coder assigned one or several codes to each participants statements, which were then discussed alongside with open or arising questions. 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 [36]. Throughout the coding process, we used analytic memos to keep track of thoughts about emerging themes. The final sets of codes are listed in App. C.

TABLE I: Inter-coder agreement of text replies. We compute Spearman correlation (*) and Cohens Kappa (†) for the most feared threat (Q1) and the replies of high/low relevance of the four investigated AML attacks. Total codes refers to the maximum number of codes given from either coder.

			Agreement	Total codes	# disagreeing	# replies
Q1	Most feared threat		.96*	232	37	136
Q8	Poisoning	high	.79†	86	6	69
		low	.65†	61	8	51
Q10	Evasion	high	.77†	67	7	54
		low	.69†	63	8	49
Q16	Membership	high	.67†	54	8	45
		low	.53†	47	11	45
Q18	M. Stealing	high	.62†	53	10	45
		low	.55†	47	9	40

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 [37], [38] for the questions about most concerning threats. This correlation coefficient, while not encompassing random overlap, allowed us to take into account how often each code was used within the single document. For the relevance coding, we instead computed Cohen's kappa [39], as we encoded high and low relevance for each of the five attacks separately, yielding several documents with varying code assignments. We report the detailed agreement measures and code numbers in Table I. We drop one attack, Q12, as it is similar to poisoning and does not yield additional insights. Given the semi-technical nature of our codebook, we consider our agreement substantial. In the following, we refer to the number of codes assigned in agreement by both coders.

V. Results

We now report the results of our study by analyzing our participants' responses and running statistical tests. The detailed statistical results including all test and sample statistics can be

TABLE II: Summary of tested statistical hypothesis and findings. Primary variables are in the columns, and are tested against secondary variables in the rows. A ✓denotes that the relationship is statistically significant, if there are brackets, some attacks are statistically significant. X means there relationship is not statistically significant and * denotes results which are not discussed in this paper.

	Estimated exposure	Threat exposure	Number of mitigations	Attack concern
Organizational Area	X*	X	X*	(✓)
AI maturity	X*	X	✓	X*
Company size	X*	X	X	X*
Techn./non techn.	X	X	X*	X
AML knowledge	X	X	✓	✓
ML knowledge	X	X	X*	(✓)
Estimated Exposure		X*	✓	(✓)
Threat Exposure	X*		✓	X

found in App. D. We analyze our sample on three layers that emerge in the context of our research questions. The first one are insights related to the occurrence of attacks in practice, the second influential factors and risks at the organizational respectively third at the individual level.

AML attacks in practice. We find evidence for AML circumventions (Sect. V-A1). Although some participants are concerned about AML and name corresponding threats explicitly (Sect. V-A2), we also find that non-AML security (e.g. access control, botnets, resource theft, etc) and privacy are encountered threats and concerns. Finally, participants name generic ML (i.e., data drift) and organizational (i.e., security awareness) challenges.

AML within organizations. In Sect. V-B1, we find no statistical significant predictors for threat exposure. At the other hand, the organization area does influence risk perception and implemented mitigation depends on exposure and concern. We also find that the reasons for threat concern are highly complex (Sect. V-B2): Our participants consider both a wider range of impacts of an attack (financial or business, up to ethics) when deciding if an attack is relevant. We find that furthermore, when an attack is perceived as irrelevant, often it is described as infeasible given deployment or use-case, or usage of the model. However, in some cases, attacks are rather seen as benign failure cases.

AML for practitioners. In Sect. V-C, we find that prior knowledge (largely in AML, but also in ML) leads to a higher concern about individual threats. Furthermore, gender sometimes influences threat perception, but not consistently.

We give a summary of all statistical results in Table II.

A. AML attacks in practice

We start with the discussion of witnessed attacks and then discuss the dreaded attacks by our participants.

1) Encountered AML threats: We first consider the estimated likelihood to become victim of an attack (Q3). Less than a fifth (17.2%) of our participants estimate the likelihood of an AML attack within the next 12 months as high or very high. Instead, roughly half (49.6%) estimate the likelihood as low or very low—indicating that exposure might not be very high. Hence, we had asked participants whether they had encountered a circumvention of their AI based workflows or systems (Q2). This was confirmed by less than a fifth (17.3%) of our participants. More concretely, seven participants (5%) witnessed one circumvention, six (4.3%) two, one (0.5%) three, two participants (1.4%) four circumventions and eight (5.7%) more than four. To obtain more in depth knowledge, we asked participants to briefly describe the circumvention in a free text field, which we now discuss.

AML in the wild. Of all replies, five (3.6%) were AML threats. Three (2.1%) described an evasion attack. The first reply was in relation to HR ("*users spam to optimize their strategy for job search*"), the second two related to autonomous vehicles ("*autonomous vehicle image recognition errors leading to dangerous path planning*"). In the case of the latter two, 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 was 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 nine replies (6.5%) contained no text, or replies like "*no details*" or "*brute force attacks*", that do not allow to deduce the exact nature of the 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 slightly more detailed: "*acquiring the data for training AI systems*"). In these cases, we assume, but cannot be sure, that they are not ML related.

Circumventions not directly related to ML. In total four (2.8%) descriptions were not ML related security threats, including resource theft (2, i.e., "*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*").

Attacks mentioned in relevance reasoning. 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, and evasion. 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 were occurrences of ML attacks in practice, namely poisoning and evasion. However, it was not always clear whether circumventions are security or safety

issues, in other words benign or attacker based failures. Furthermore, almost a third of our participants' replies remained vague, not allowing to understand the exact nature of an attack. Almost another third of replies were data breaches, privacy leaks, or other non-AML security issues.

2) Concerns about AML: We further aimed to understand what AML challenges practitioners face (Q1). To avoid priming, we had asked this before mentioning any specific AML attack. Of all participants, almost all (93.5%) provided a reply, and more than a fifth (22.9%) provided more than one concern. In the following text, as more than one code could be assigned to a reply and the total number of assigned codes is not equal to the number of participants, we report no percentages.

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 challenge to be compliant with legislation (6). Several (20) challenges concerned security in organizations. Corresponding replies are for example "*convincing stakeholders of the risks*", "*protecting intellectual 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 non-AML security, 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) was related to data. While some of these replies were vague (11, "*data leak*"; 17, "*data security*"), some were related to 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 through data (through training for poisoning, through test data for evasion, membership inference and model stealing). However, threats caused by data could also include non-AML security, data quality, privacy, etc. We thus leave more detailed research on this question for future work.

Conclusion. There were few, but some concrete mentions of AML. Although we had explicitly asked for AML, participants also raised non-AML security and privacy concerns, reasoning that these were more pressing than AML. Concerns also encompassed organizational challenges related to ML itself or risk communication or assessment. Finally, participants often reasoned vaguely about data security, leaving open whether they refer to data quality, privacy, or AML issues.

3) Conclusion: We find that AML threats did occur in practice, and that some participants were explicitly concerned about AML. At the same time, it remains sometimes (for example in evasion) unclear whether an incident is a security or safety issue. Furthermore, concerns encompass non-AML security, privacy, organizational challenges and ML problems such as dataset shift, for example.

B. AML within organizations

In this subsection, we examine AML in practice from an organizational perspective. To this end, we first relate different questions via statistical tests, and then analyze the arguments about individual attack relevance from our participants.

1) Organizations approaches to ML security: We first analyze whether the organization area influences threat perception, and then attempt (but fail) to find factors from our questionnaire that predict threat exposure. Finally, we investigate which factors influence the implementation of mitigations.

Organization area and threat perception. We assumed that the area an organization operates in affects threat perception. For example healthcare is based on sensitive data, thus healthcare workers may be more concerned about related threats, in this case membership inference. Our sample contains two large industry groups (Q25): IT security (15 participants) and healthcare (18 participants). We tested for both groups whether there were threats perceived as more relevant compared to the rest of the sample. We divided the sample into one subgroup fulfilling the criteria and the rest of the sample, and used a Mann-Whitney-U test to determine if concern deviated in a statistical significant manner. For healthcare, we investigated membership inference but found no statistical significance ($p = 0.7$). In the case of security companies, the relevance of backdoor attacks was statistically significant ($p = 0.002$), as well as evasion ($p = 0.02$) and membership inference ($p = 0.026$). Other threats did not exhibit statistical significance.

Predicting threat exposure. When it comes to threat exposure, we assumed that both the organization area plays a role as well as the amount of exposed AI technology of the organization and its visibility. In our questionnaire, these were the organization area as defined in the previous paragraph, AI maturity (Q24) and organization size (Q18). The organization area was tested with a Mann-Whitney-U test, the latter two with an ordinal regression model. We did not find any statistically significant relations.

Predicting threat concern. We expected that the individual threat concern may depend on both threat exposure (Q2) and estimated exposure (Q3), and thus tested both with an ordinal regression model. In terms of threat exposure, we could find no statistical significant relationship to any attack. In case of estimated exposure, we do find that it statistically significantly predicts both concern in case of poisoning ($p = 0.002$) and evasion ($p = 0.003$).

Predicting the amount of implemented mitigations. We asked our participants about the number of implemented mitigations (Q4, for example approaches like "*documentation*", "*fail safe plans*", or "*incident response*"). We assumed that the implementation of these mitigations depends on factors such as previous exposure to threats, estimated risk to become victim of an attack, organization size (how much personnel can be dedicated to securing models), or how long models are in production. More concretely, we tested if the number of implemented mitigations (0-7) was influenced by factors like exposure to threats (Q2), estimated risk to become a target of an AI circumvention (Q3), organization size (Q18), and AI maturity (Q25). We used an ordinal regression to model these relationships and found that exposure ($p = 0.012$), estimated risk ($p = 0.013$) and AI maturity ($p = 0.004$) were statistically significant predictors.

Conclusion. We found that while the organization area affects the perception of some threats significantly, there were no statistically significant variables for threat exposure. The amount of implemented mitigations was however statistically related to threat exposure, estimated risk, and AI maturity.

2) 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 studied factors on threat concern such as the ease to attack and defend, or possible benefit of carrying out the attack [17]. We instead asked our participants without priming to give a short reason for the relevance or irrelevance of an AML threat given a two sentence description (but not the name) of the attack. More concretely, we asked our participants how relevant they thought poisoning (Q8), evasion (Q10), membership inference (Q16) and model stealing (Q18) was. In this version of the paper, we do not discuss backdoors (Q12), as they are similar to poisoning. 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.*", Q14). Although some participants reported high concern, the threat was rated statistically significantly less relevant compared to all other threats[5]. We thus omit the sanity threat in the following discussion, where we first discuss the high relevance of each poisoning, evasion, membership inference, and model stealing. The same order is used for the discussion of irrelevance replies. A summary of our results is depicted in Table III, and we plot the numerical relevance ratings in Figure 4.

Poisoning–high relevance. The most frequently coded reply reasoning for relevance, occurring 10 times, was the relevance within the applications setting of the participant ("*we use AI for security purposes, tampered training data is one of the best ways for attackers to evade the system*"). Following up codes were 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 organization 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*").

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*").

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 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*").

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*").

[5]Mann-Whitney-U test with $[1.4e^{-10} < p < 1.2e^{-16}]$.

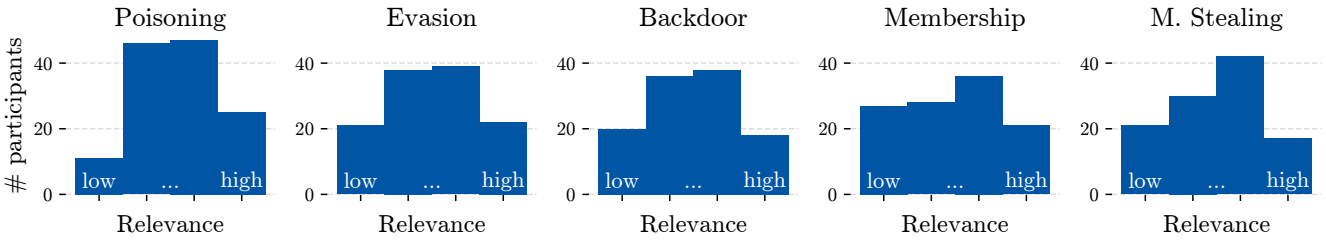


Fig. 4: Reported relevance for the all five attacks we tested. The Likert-scale provided had four items ranging from "*irrelevant*" to "*very relevant.*"

Poisoning–low relevance. The most frequent code (14) for irrelevance of poisoning attacks was that the data was not accessible to 3rd parties or the outside of the organization ("*no one can access the training samples*"). Additional frequent codes were that the threat is not relevant under the 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 stated that their human in loop (5, "*the training data is curated by us*") or another defense they implemented (5, "*very few publicly available data used for training*") prevented the attack. Finally, some (3) also reason that the attack is hard to carry out.

Evasion–low relevance. Most participants (11) arguing against the relevance of 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*").

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*").

Model stealing–low relevance. Most participants (13) that deemed model stealing irrelevant reasoned based on their use-case ("*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*"). Otherwise, participants remarked that their models were not accessible (7, "*we don't offer API's to our models.*"), or were replaced often and copying them yielded no benefit (3, "*model is continuously updated, and previous models don't have much value*"). Participants also reasoned that the attack was hard to carry out (4), or generally irrelevant (3, "*this is a business model issue, not a technical issue*"). Participants also doubted that an attacker might benefit (3, "*the value of copying [our models] would be quite small for someone else*"), or reasoned that the attack does not apply in their use case or way to deploy ML (3, "*the model is likely to be deployed on edge devices so it will be anyway known to the potential attacker.*").

TABLE III: Participants' argumentation for the relevance of attacks. For each attack we present the five most frequent (ties broken randomly) arguments (and their frequency).

	RELEVANCE	IRRELEVANCE
Poisoning	Relevant in application (10)	(14) No data access
	Impact on safety (9)	(9) Not relevant in application
	Impact on performance (9)	(8) Doubting attacker
	Relevance without argument (9)	(5) Human in the loop defense
	Hard to defend (7)	(5) Some defense implemented
Evasion	Impact on model performance (11)	(11) No data access
	Impact on safety (6)	(10) Not relevant in use case
	Impact on decision making (4)	(7) Doubting attacker
	Easy to do (4)	(4) Some defense implemented
	Hard to defend (4)	(3) Hard to do
Membership	Data breach (21)	(10) Not relevant in use case
	Relevance without argument (3)	(9) No sensitive data
	Business information leakage (3)	(4) Irrelevance without argument
	Regulatory compliance (3)	(3) No query access
	Relevant in application (3)	(3) Doubting attacker
M. Stealing	Impact on intellectual property (8)	(13) Not relevant in use case
	General business impact (5)	(7) No query access
	Profit for competitor (5)	(4) Hard to do
	Financial loss (4)	(3) Doubting attacker
	Attacker credible (3)	(3) Model shortlived

Conclusion. Our analysis shed light on the complexity of AML in practice. While poisoning was the only attack to be witnessed in practice by several participants, other threats were deemed relevant due to their (potential) impact. Such impact was very diverse, and ranged from decreased model performance, wrong decision making, to business implications like leak of information, financial loss and leakage of intel-

lectual property. When an attack was deemed irrelevant, the attacker often would not have access to the required data. In this sense, both application and deployment are orthogonal factors influencing vulnerability: One use-case may be security critical only when deployed in a certain way, otherwise not. Finally, sometimes the difference between security and safety was not well distinguished.

3) Conclusion: Practitioners from IT security companies were significantly more concerned about AI threats. At the same time, in our sample, exposure to threats was not a factor of organization size or AI maturity. The amount of implemented mitigations did, on the other hand, depend on previous threat exposure, but also on expected risk, and AI maturity. Our participants, when reasoning about attack relevance, encompassed not only feasibility or the ease to mitigate an attack, but considered direct impacts like financial loss, information leakage, or business harm. Furthermore, decision making based on ML or practical encounters fuel concern. When threats were deemed irrelevant, this was usually based on specifics of the use-case or deployment and/or inaccessibility or the required resources or data for the attacker. Finally, in some cases, attacks were perceived as safety issues, or in other words benign failure cases.

C. AML for practitioners

In this section, we investigate individual factors that may influence threat exposure, estimated threat exposure or threat perception. Such individual factors are the role in team, prior knowledge in AML, ML, prior education in general, or gender.

Role in team. Assuming that an ML engineer is closer to model deployment than a product manager, we may assume that technical roles in teams are more exposed to threats. We thus investigated the reported role within a team (Q30) in relation to threat exposure (Q2), expected exposure, and perceived attack relevance. To this end, we split our sample into technical (48 participants, for example "*ML Engineer*", "*ML Scientist*", "*data architect*", etc) and non-technical (91, for example "*product owner*", "*auditor*","*domain expert*", etc). We found, using a Mann-Whitney-U test, that threat exposure was not different for technical and non-technical roles ($p = 0.15$), and there was no difference in expected exposure ($p = 0.8$). Furthermore, we tested using a Mann Whitney-U test whether the concern of these two groups differed statistically. For no threat, this was the case ($0.4 < p < 0.8$).

Prior knowledge in AML. Another possible individual factor is knowledge of AML. Knowledge means understanding, thus potentially raising threat concern and also motivating countermeasures or mitigations. In Q32, we asked our participants to self-report their knowledge in AML. We split our sample into two groups (knowledgeable 40, not knowledgeable 99) and tested for statistical significance in concern about threats, exposure to threats (Q2) and the number of mititgations implemented (Q32) using the Mann-Whitney-U test. There was no statistically significant difference between these two groups for threat exposure ($p = 0.15$) or general threat concern ($p = 0.5$). However, participants who reported knowledge of AML were significantly more concerned about AML threats (except sanity, $0.006 < p < 0.018$). We found finally that self-reported prior knowledge lead to a statistically significantly increased number of implemented mitigations ($p = 0.0024$).

ML knowledge and education. Given the high relevance of AML knowledge, we also tested the influence of general ML knowledge and education (e.g., Highschool, Bachelor, PhD) on both exposure (Q2), expected exposure (Q3), and threat concern. While we expected ML knowledge to have an influence on exposure and threat concern (most AML attacks are based on ML mechanisms), we did not expect general education to influence concern. We tested our hypotheses with an ordinal regression model, and found no statistical significance for exposure. For the individual threats, we found that only ML knowledge significantly influenced concern for evasion ($p = 0.025$) and membership inference ($p = 0.025$), but for no other threat ($p > 0.5$).

Gender. We further investigated whether gender influences expected threat exposure or threat concern. We expected that women were similarly or more concerned, in line with previous findings [40]. To test this hypothesis, we divided our sample into female (20 participants) and male (99 participants) and computed a Man-Whitney-U test on the replies for estimated exposure (Q3) and threat concern. Gender did not statistically influence the overall estimated risk ($p = 0.43$). For most threats, there was no significant difference ($p > 0.15$) either. However, for model stealing ($p = 0.021$), we found that women are less concerned, contrary to our expectations. A possible explanation could be that women work with different applications and deployment settings, but our data did not allow to investigate this profoundly.

Conclusion. In this subsection, we investigated factors specific to individual practitioners and their relation to threat exposure or perception. We find that prior knowledge (largely in AML, but also in ML) lead to a higher concern about individual threats. We furthermore find that gender influences threat perception, too, but not consistently.

VI. Limitations

In this section, we discuss limitations in our study that affect the generalizability of our results. We first describe limitations within the sample, then within the questionnaire, and finally within the statistical approach.

Sample limitations. Our sample is limited to English speaking practitioners, and biased towards the global north. Furthermore, we had initially planned to compare opinions from industry and academia, but got feedback early on that our questionnaire was too industry specific for academics. More specifically, some participating academics reported back to us to have filled the survey from the perspective of recent industry experience. In contrast to our expectations, the industry area does not allow to deduce which participants are from academia, making it hard to understand the influence academics could have had on the results. Due to the usage of different links to monitor recruiting strategies, however, we know that less than a quarter (35) of participants are pure academics without industry experience. Still, we might underestimate the occurrence of threats in the wild (Sect. V-A1).

Questionnaire limitations. We did only consider self reported knowledge, and did not asses our participants' knowledge. Independently, we failed to observe statistically significant results for type of learning (supervised, unsupervised), kind of labels (none, categorical, real), and type of data (vision, video, etc.). A possible cause for this, apart from the conclusion of indeed no relation, is that participants for example work in several projects. This diversity was not foreseen by us when designing the questionnaire.

Methodological limitations. We performed a sample size estimation upfront before designing the questionnaire (see Sect. III). However, the Mann-Whitney-U test's sample size is dependent on factors such as mean and variances of both samples, factors that are impossible to approximate upfront. While in most cases, differences indicate that our sample size is enough, we cannot exclude that in some cases, the test returns a too conservative result (e.g., masking significance when there is indeed an effect in the data). These cases were when testing threat concerns for the role in team (Sect. V-C), and when testing whether healthcare workers are more concerned about threats like membership inference (Sect. V-B1). These effects should be re-evaluated in future work with a different design or more participants.

VII. Implications and Future Work

Despite these limitations, our research yields practical implications for a better understanding of attack relevance, more granular risk assessments and an improved communication of AML risks. We now concretise these implications and conclude the paper by discussing open research questions that could be addressed by future work.

Better understanding of general attack relevance. For each of the attacks we tested for, participants' argumentation for relevance involved reasoning about the validity of a threat in a certain application, deployment, or use case (Table III). More specifically, not all attacks apply in all applications or forms of deployment. This finding is highly relevant for risk management in the real-world and implies that threat modelling should always consider the specific context of an AI system. To this regard, further research is needed to investigate which attacks should be considered in which application scenarios, and which deployment is prevalent in which application. Somewhat orthogonal, a quantification of the impact incurred by an attack (in terms of financial loss, for example) based on attack type and application would benefit a deeper understanding of risks related to ML in the wild.

More granular risk assessments of AI systems. Our findings in Section V-B help to understand which factors should be taken into account when threat modelling an AI application. This is relevant for risk assessments of real-world applications. According to international standards for information security management such as ISO/IEC 27001, these assessments should evaluate the likelihood of threats and the potential impact if they materialize. Our code book that evolved based on participants' statements on attack relevance confirms this approach (Table V). In addition, the concrete codes for 'relevance' and 'impact' that we found in developing our codebook can be used as an orientation for practitioners in trying to concretize likelihood and impact within risk assessments of AI systems. For example, we show that an AI auditor should evaluate concrete implications of 'financial loss' or 'business information leakage' in order to assess the potential impact of an AI risk that might materialize. Thus, our findings allow risk assessments of AI systems to become more granular.

Improved communication of AML risks. We provide insights into why practitioners think specific attacks are relevant or irrelevant (Sect. V-B2). These insights into the rationale of relevance for attacks could be a starting point for educational measures to increase AML awareness in organizations that deploy AI. More concretely, our results could help educating business stakeholders that they, for example, have to hedge against model stealing because it is a potential target for IP theft, or that they should consider the tangible risk of poisoning as it may affect their decision-making regarding technology setups or engineering processes. This unveiling of rationales behind attacks might ease the communication of AML risks.

Open research questions. As we have seen in the previous section, our study comes with some limitations that can be overcome in future work. For example the dimensions knowledge, role in team and the application area deserve to be studied more in depth. Along these lines, we found that not only the application, but also the deployment is a crucial factor determining the vulnerability of an ML system (Sect. V-B2). Both factors need to be monitored and can then jointly, for example together with 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 the deployment setting, for example.

Also, some aspects of the relationship between knowledge and concern about threats (Sect. V-C) remain unclear. This relationship is similar to a chicken-egg problem: more knowledge might imply more sensitivity towards threats, but at the same time more concern also brings about the need for more information, (hopefully) leading to the acquisition of more knowledge. A clear understanding of cause and effect here would benefit regulation and AI in practice.

VIII. Related Work

In this section, we first put our findings in relation to other surveys that address AML in practice. Afterwards, we discuss works that are overall less related, yet provide important findings in relation to our insights.

As there is no other survey with a similar amount of participants and thus depth and statistical findings, we instead discuss specific findings from previous works that our larger sample either confirms or contradicts. For example, Kumar et al. [16] found that poisoning is the most feared AI security threat by companies, a finding we confirm (Fig. 4). Our findings also support, analogous to their reports, that practitioners are still very much concerned with traditional security (Sect. V-A2) [16]. Kumar et al. [16] furthermore state that AML is generally perceived as "futuristic". We can not confirm

this, however have to emphasize that there are 2-3 years between our surveys. Boenisch et al. [19] find that their ML-security score is overall low, but is increased when participants are developing, and not only applying, ML. We can not confirm these results (Sect. V-C), however have to emphasize that while we measure direct concern for one ML attack or concern of AML attacks in general, Boenisch et al.'s [19] score entails also cybersecurity, and ML in general.They [19] furthermore study privacy, which we only consider in the sense of membership inference. Mirsky et al. [17] also conducted a survey that aims to understand offensive AI. Specifically for AML, they asked participants how profitable, harmful, detectable and achievable attacks are. We go a step further in our survey and allow our subjects to freely reason why (or why not) they believe an attack to be relevant (Sect. V-B2). Finally, Mirsky et al. [17] expect offensive AI techniques to manifest within the next 12 months, a finding our sample does not support (Sect. V-A1).

More loosely related is the work by Bieringer et al. [18] who conducted semi-structured interviews with industrial AI practitioners. They find that malicious ML circumventions already take place in industry, which we confirm (Sect. V-A1). Our study answers their question about the importance of education in the affirmative (Sect. V-C). Finally, we investigated the influence of gender on (ML) security perception, as prior works found that women are overall more concerned about non-AML security [40]. While we do not find an overall difference, some specific attacks are rated differently (Sect. V-C), possibly a consequence of women working in slightly different applications [41].

Finally, our finding of conflated security and safety concepts (Sect. V-A2 and Sect. V-B2) has been reported in the cybersecurity domain: For example, Gross and Rosson [42] found in their study that end users do not distinguish system failure from external attacks, and reason that this is a valid level of abstraction for consumers. Moreover, early work from 1992 in non-AML security by Loch et al. [43] showed that managers, when introducing information systems to their organizations, ranked intentional security lower than safety. Both works, albeit carried out on different populations than our study, help contextualise our findings (Sect. V-A2 and Sect. V-B2): a similar gap between tomorrow's reality and today's understanding might also apply for AML.

IX. Conclusion

To overcome the lack of knowledge on AML in practice, we conducted a survey of 139 industrial practitioners opinions on ML security and attack relevance. We found evidence for AML attacks, more specifically evasion and poisoning, in practice. However, it remains often unclear whether an incident is a security or safety issue. In addition, also privacy, ML, and organizational challenges like data drift are of importance to our participants. In terms of the organizational aspects of AML we find that companies from some industry areas like IT security are more concerned about some ML attacks. We find no variable that is statistically related to threat exposure. Exposure, together with expected risk and AI maturity, however predicts the amount of implemented mitigations. 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. Finally, on the individual level, we find that self reported knowledge, in particular in AML, increases attack concern. Our results yield important insights for regulators and auditors as we analyze relevance and irrelevance, and point out that the boundary between safety and security is not always clear. We are further confident that we are contributing towards more research eliciting when ML systems are vulnerable and which factors influence vulnerability and threat perception.

Acknowledgments

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 supporting our recruitment. The research reported in this paper has been partly funded by BMK, BMDW, and the Province of Upper Austria in the frame of the COMET Programme managed by FFG in the COMET Module S3AI; and by Fondazione di Sardegna under the project "TrustML: Towards Machine Learning that Humans Can Trust", CUP: F73C22001320007.

References

[1] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *CCS*, 2006, pp. 16–25.
[2] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Patt. Rec.*, vol. 84, pp. 317–331, 2018.
[3] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv*, 2017.
[4] A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, and F. Roli, "Wild patterns reloaded: A survey of machine learning security against training data poisoning," *arXiv*, 2022.
[5] N. Dalvi, P. Domingos, Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *KDD*, 2004, pp. 99–108.
[6] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the ML model supply chain," *arXiv*, 2017.
[7] Y. Ji, X. Zhang, and T. Wang, "Backdoor attacks against learning systems," in *IEEE CNS*, 2017, pp. 1–9.
[8] S. J. Oh, B. Schiele, and M. Fritz, "Towards reverse-engineering black-box neural networks," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 2019, pp. 121–144.
[9] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv:1605.07277*, 2016.
[10] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014.
[11] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *USENIX Sec.*, 2016, pp. 601–618.
[12] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *S&P*, 2017, pp. 3–18.
[13] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *NeurIPS*, vol. 33, pp. 1633–1645, 2020.
[14] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *S&P*, 2010, pp. 305–316.
[15] J. Gilmer, R. P. Adams, I. Goodfellow, D. Andersen, and G. E. Dahl, "Motivating the rules of the game for adversarial example research," *arXiv*, 2018.
[16] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia, "Adversarial machine learning-industry perspectives," in *S&P Workshops)*, 2020, pp. 69–75.

[17] Y. Mirsky, A. Demontis, J. Kotak, R. Shankar, D. Gelei, L. Yang, X. Zhang, W. Lee, Y. Elovici, and B. Biggio, "The threat of offensive ai to organizations," *arXiv*, 2021.
[18] L. Bieringer, K. Grosse, M. Backes, B. Biggio, and K. Krombholz, "Industrial practitioners' mental models of adversarial machine learning," in *SOUPS*, 2022, pp. 97–116.
[19] F. Boenisch, V. Battis, N. Buchmann, and M. Poikela, ""i never thought about securing my machine learning systems": A study of security and privacy awareness of machine learning practitioners," in *Mensch und Computer*, 2021, pp. 520–546.
[20] P. J. Muller, S. E. Young, and M. N. Vogt, "Personal rapid transit safety and security on university campus," *Transportation research record*, vol. 2006, no. 1, pp. 95–103, 2007.
[21] B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, and J. D. Tygar, "Antidote: understanding and defending against poisoning of anomaly detectors," in *IMC*, 2009, pp. 1–14.
[22] B. Biggio, B. Nelson, and P. Laskov, "Support vector machines under adversarial label noise," in *ACML*, 2011, pp. 97–112.
[23] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *CCS*, 2018.
[24] M. Juuti, S. Szyller, S. Marchal, and N. Asokan, "Prada: protecting against dnn model stealing attacks," in *EuroS&P*, 2019.
[25] H.-Y. Lin and B. Biggio, "Adversarial machine learning: Attacks from laboratories to the real world," *IEEE Comp.*, vol. 54, no. 5, pp. 56–60, 2021.
[26] N. Nachar *et al.*, "The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution," *Tutorials in quantitative Methods for Psychology*, vol. 4, no. 1, pp. 13–20, 2008.
[27] R. Ferber, "Order bias in a mail survey," *Journal of Marketing*, vol. 17, no. 2, pp. 171–178, 1952.
[28] N. Huaman, B. von Skarczinski, C. Stransky, D. Wermke, Y. Acar, A. Dreißigacker, and S. Fahl, "A large-scale interview study on information security in and attacks against small and medium-sized enterprises," in *USENIX Security*, 2021.
[29] K. Mandia and C. Prosise, *Incident response: investigating computer crime*. McGraw-Hill, Inc., 2001.
[30] E. U. Commission *et al.*, "Commission recommendation of 6 may 2003 concerning the definition of micro, small and medium-sized enterprises," *official Journal of the EU*, vol. 46, no. L124, pp. 36–41, 2003.
[31] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *arXiv*, 2017.
[32] R. B. Miller, "Response time in man-computer conversational transactions," in *FJCC*, 1968, pp. 267–277.
[33] H. Suresh, S. R. Gomez, K. K. Nam, and A. Satyanarayan, "Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs," in *CHI*, 2021.
[34] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser *et al.*, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
[35] Kaggle. (2021) State of machine learning and data science. [Online]. Available: https://storage.googleapis.com/kaggle-media/surveys/Kaggle's%20State%20of%20Machine%20Learning%20and%20Data%20Science%202021.pdf
[36] A. Strauss and J. Corbin, *Basics of qualitative research*. Sage publications, 1990.
[37] N. McDonald, S. Schoenebeck, and A. Forte, "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice," *ACM on HCI*, vol. 3, no. CSCW, pp. 1–23, 2019.
[38] L. Jinyuan, T. Wan, C. Guanqin, L. Yin, F. Changyong *et al.*, "Correlation and agreement: overview and clarification of competing concepts and measures," *Shanghai Archives Psych.*, vol. 28, no. 2, p. 115, 2016.
[39] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.
[40] M. A. Hossain, "Security perception in the adoption of mobile payment and the moderating effect of gender," *PSU Research Review*, 2019.
[41] M. M. Marini, P.-L. Fan, E. Finley, and A. M. Beutel, "Gender and job values," *Sociology of Education*, pp. 49–65, 1996.
[42] J. B. Gross and M. B. Rosson, "Looking for trouble: understanding end-user security management," in *CHIMIT*, 2007, pp. 10–es.
[43] K. D. Loch, H. H. Carr, and M. E. Warkentin, "Threats to information systems: today's reality, yesterday's understanding," *Mis Quarterly*, pp. 173–186, 1992.

APPENDIX A
QUESTIONNAIRE

Part I - Security of AI within your organization

Q1: In your daily work and your organization's AI workflows, products or systems - what are the most pressing security challenges? [text field]
Q2: Did you already experience a circumvention of your AI based workflows, products or systems? [yes/no]
IF YES: **Q2.1**: How many circumventions of your AI based workflows, products or systems have you experienced? [1,2,3,4,>4]
Q2.2: Please describe the most severe circumvention of your AI based workflows, products or systems. [text field]
Q3: How high do you estimate the risk of becoming a victim of an attack related to your AI based workflows, products, or systems within the next 12 months? [1 (very low) to 5 (very high)]
Q4: Which of the following approaches does your organization implement in terms of the security of your AI based workflows, products, or systems? [None, Documentation, Guidelines, Mitigations, Fail safe plans, human in the loop, incident response, security testing, other]

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.
Q5: 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]
Q6: Why do you think this threat scenario is (placeholder for previous selection)? [text field]
These 2 questions are repeated iteratively all attacks:

a) **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.
b) **9,10**: 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.
c) **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.
d) **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.
e) **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.
f) **17,18**: Given an API / black box access to a model, copy its functionality. In other words, repeatedly observe

TABLE IV: 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	NonAMLSec	General	Privacy	General	ML	General
	Poisoning		Libraries		Regulations		Explainability
	Evasion		Access				Bias
	Model Stealing		CustomerIsRisk	Data	General		Concept drift
	Performance impact		CodeBreach		Data sharing		
	Robustness		3rdParty Provider		Breach	Organization	Complexity
	TestTime		Precise threat		Sensitive data		IP
	TrainingTime		Cloud		Classify if sensitive		TradeOffs
	ModelItself						SecurityAwareness
							Human Harm

TABLE V: Codes used for the attack relevance, where participants argue why (or why not) 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				Externalization of responsibility
	Other threat more likely				
	Safety				

input and output pairs from the model to reproduce its functionality.

Part II - AI within your organization

Q17: In which country is your organization headquartered? [drop down with all countries]

Q18: What is the number of employees at your organization? [<10, 10-49, 50–99, 100–249, 250–499, >500]

Q19: 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]

Q20: What kind of data analysis do you work with primarily? [supervised learning, unsupervised learning, semi-supervised learning, reinforcement learning, other]

Q21: What do you use AI for primarily (e.g. sentiment analysis, object detection, malware classification)? [text field]

Q22: 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:]

Q23: What kind of labels do you work with primarily? [unlabelled, categorical, real valued, structured data, other]

Q24: 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

Q25: Which of these goals are part of your organization's ML-model checklist? [Performance, fairness, explainability, security, privacy, ethics, system response, other]

Part III - Demographics and your AI background

Q26: In which year were you born? [1935-2021]

Q27: What gender do you identify with? [Female, male, other, I do not want to disclose]

Q28: In which country are you located? [drop down with all countries]

Q29: What is your level of education? Please specify the highest. [Highschool, Bachelor, Master/Diploma, Training/Apprenticeship, PhD, Other]

Q30: What is your role in your team? [ML Engineer, ML researcher, Data scientist, Domain Expert, Product Owner, Auditor, Other]

Q31: Please complete the following sentence. When it comes to machine learning, I believe I have... [No knowledge, a little/some/moderate/high knowledge]

Q32: In which of these areas have you taken a lecture or intense course? [None, Machine learning, Security, Adversarial Machine Learning]

APPENDIX B
DETAILED COMPARISON WITH THE KAGGLE SAMPLE

As we write in Sect. IV-A in the sample description, our sample matches roughly the numbers from the Kaggle report [35], modulo that the report has a larger sample (¿25,000 participants) and has made slightly different design choices concerning the questions. In this Appendix, we redraw and reordered the plots from the report to be able to roughly compare them to our data in table VI.

APPENDIX C
COMPLETE SETS OF CODES

We here depict the full sets of codes for most feared threat/the first question in Table IV. The codes for the attack relevance coding are listed in Table V.

APPENDIX D
DETAILED RESULTS OF STATISTICAL TESTS

We here report the detailed results for all statistical tests in the main paper in the order of appearance or mentioning.

A. *Statistical tests from Section V-B1*

We first review the tests from Section V-B1 about the participant's organization in the order of the paragraphs in the main paper.

Organization Area and threat perception. In this part, we depict the detailed results of the organization area and threat relevance ratings. The table takes organizational area (Q19), and splits according to Healthcare / Security the provided rating (e.g., participants working in a healthcare setting vs all other participants). We depict mean (± standard deviation) and the sample size for each subgroup (left/sample I: works in specified industry, right/sample II: remaining participants) as well as the test statistic and the p-value.

Sample I		Sample II			
μ (±sdt.)	#	μ (±sdt.)	#	U	p
Healthcare: Membership Rating (Q13)					
1.67 (± 1.49)	18	1.71 (± 1.85)	121	1030.0	0.71
Security: Poisoning Rating (Q5)					
3.0 (± 0.97)	15	2.3 (± 1.38)	124	1200.0	0.06
Security: Evasion Rating (Q7)					
2.73 (± 1.44)	15	1.9 (± 1.61)	124	1262.0	**0.02**
Security: Backdoor Rating (Q9)					
3.0 (± 0.97)	15	1.6 (± 1.74)	124	1372.5	**0.0**
Security: Sanity Rating (Q11)					
0.4 (± 2.06)	15	0.24 (± 1.82)	124	953.0	0.88
Security: Membership Rating (Q13)					
2.6 (± 1.5)	15	1.6 (± 1.81)	124	1251.5	**0.03**
Security: M. Stealing Rating (Q15)					
2.13 (± 1.5)	15	1.65 (± 1.81)	124	1047.0	0.42

Predicting threat exposure. Analogous to the previous tests, we again divide according to company area (Q19) and now test for exposure.

Sample I		Sample II			
μ (±sdt.)	#	μ (±sdt.)	#	U	p
Security: exposure (Q2)					
0.44 (± 1.21)	18	0.51 (± 1.32)	121	1079.0	0.93
Healthcare: exposure (Q2)					
1.0 (± 1.86)	15	0.44 (± 1.21)	124	1038.5	0.27

We furthermore ran a regression model optimized with bfgs, with the predictors AI maturity (Q24) and company size (Q18). The obtained log-likelihood of the model was -96.907, the AIC 207.8 and the BIC 228.4. The model had 132 residuals and 7 degrees of freedom.

	coef	std err	z	P> |z|	[0.025	0.975]
AI maturity	0.1	0.074	1.33	0.19	-0.05	0.24
Comp. size	0.01	0.061	0.13	0.9	-0.11	0.13
0.0/1.0	1.36	0.374	3.64	0.0	0.63	2.09
1.0/2.0	-1.5	0.364	-4.11	0.0	-2.21	-0.78
2.0/3.0	-1.38	0.393	-3.51	0.0	-2.15	-0.61
3.0/4.0	-2.97	0.991	-3	0.0	-4.91	-1.03
4.0/5.0	-2.15	0.695	-3.1	0.0	-3.52	-0.79

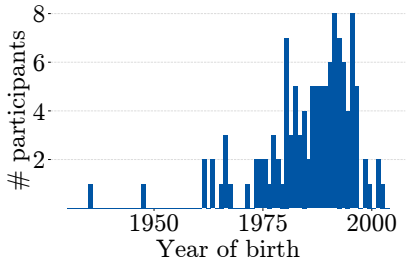
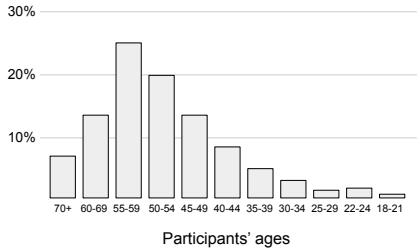
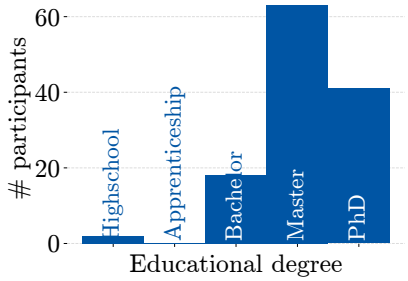
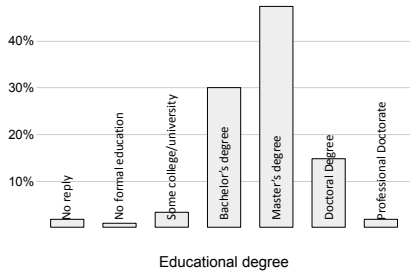
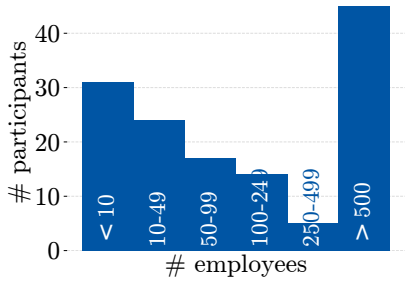
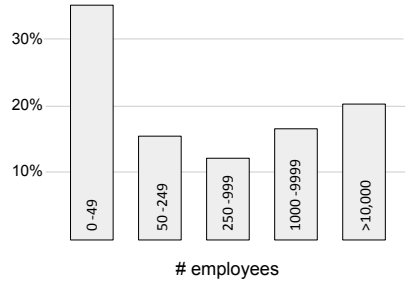
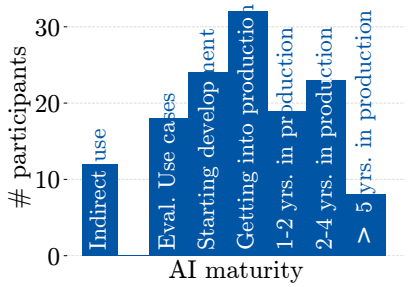
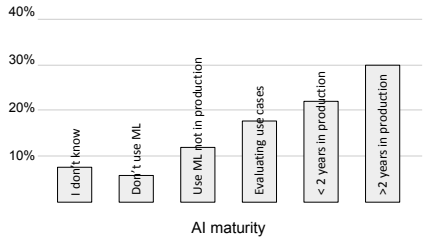
Predicting amount of implemented mitigations. To gain insights on the amount of implemented mitigations, we ran a regression model optimized with bfgs. The predictors were AI maturity (Q24), company size (Q18), exposure (Q2) and estimated risk of exposure (Q3). As predicted variable, we determine the number of implemented mitigations (min. 0, maximum 7). The obtained log-likelihood was -266.99, the AIC 556 and the BIC 588.3. The model had 128 residuals and 11 degrees of freedom.

	coef	std err	z	P> |z|	[0.025	0.975]
AI maturity	0.15	0.051	2.84	**0.00**	0.045	0.246
Comp. size	0.01	0.043	0.17	0.87	-0.08	0.09
Exposure	0.18	0.071	2.5	**0.01**	0.04	0.32
Est. risk	0.2	0.079	2.5	**0.01**	0.04	0.35
0.0/1.0	-0.06	0.316	-0.18	0.86	-0.68	0.56
1.0/2.0	-0.57	0.211	-2.7	0.01	-0.98	-0.15
2.0/3.0	-0.57	0.179	-3.2	0.00	-0.92	-0.22
3.0/4.0	-0.68	0.180	-3.8	0.00	-1.03	-0.32
4.0/5.0	-0.91	0.220	-4.1	0.00	-1.34	-0.47
5.0/6.0	-1.02	0.273	-3.7	0.00	-1.56	-0.49
6.0/7.0	-1.24	0.362	-3.4	0.00	-1.95	-0.53

Comparing attack ratings. As reported in the main paper, the Sanity threat (see App. A) is rated statistically significantly different than the other attacks. We here report the detailed results from the attack vs attack ratings. Here, we use the two samples with the numeric inputs for one attack as one input sample. The statistics (μ,std., number of samples) are reported for each attack. we encode replies without rating as well.

In addition to the statistical significance for sanity, we also observe statistical significance for poisoning. We do not report this, as we were not able to randomize the order of the attacks, and we assume this is an effect of our participants getting tired. An alternative explanation is that as reported by Kumar

TABLE VI: Detailed comparison with the Kaggle report [35]. We reordered the original information to enable an easier comparison. The AI maturity data is from the Kaggle report from 2020, as the report from 2021 does not contain this information. **Units of plots do not match and have not been adjusted.**

	Our sample	Kaggle [35]
Gender	70.1% male 14.3% female 7% do not wish to disclose 9% no reply	82.2% male 16.2% female 0.2% non-binary 1.2% prefer not to say 0.2% prefer to self describe
Age		
Education		
Organizations size		
AI maturity		

et al. [16], poisoning is indeed the most feared threat (and thus rated higher then other attacks). In addition, the test statistic is also higher when testing against santity for any attack, showing that this effect is stronger.

Attack I		Attack II			
μ (±sdt.)	#	μ (±sdt.)	#	U	p
Poisoning (Q5) 2.37 (± 1.36)	139	Evasion (Q7) 1.99 (± 1.61)	139	10908	0.05
Poisoning (Q5) 2.37 (± 1.36)	139	Backdoor (Q9) 1.76 (± 1.73)	139	11557	**0.0**
Poisoning (Q5) 2.37 (± 1.36)	139	Sanity (Q11) 0.26 (± 1.85)	139	15796	**0.0**
Poisoning (Q5) 2.37 (± 1.36)	139	Membership (Q13) 1.71 (± 1.81)	139	11622	**0.0**
Poisoning (Q5) 2.37 (± 1.36)	139	M. Stealing (Q15) 1.71 (± 1.79)	139	11566	**0.0**
Evasion (Q7) 1.99 (± 1.61)	139	Backdoor (Q9) 1.76 (± 1.73)	139	10327	0.31
Evasion (Q7) 1.99 (± 1.61)	139	Sanity (Q11) 0.26 (± 1.85)	139	14669	**0.0**
Evasion (Q7) 1.99 (± 1.61)	139	Membership (Q13) 1.71 (± 1.81)	139	10441	0.23
Evasion (Q7) 1.99 (± 1.61)	139	M. Stealing (Q15) 1.71 (± 1.79)	139	10383	0.27
Backdoor (Q9) 1.76 (± 1.73)	139	Sanity (Q11) 0.26 (± 1.85)	139	14065	**0.0**
Backdoor (Q9) 1.76 (± 1.73)	139	Membership (Q13) 1.71 (± 1.81)	139	9789	0.84
Backdoor (Q9) 1.76 (± 1.73)	139	M. Stealing (Q15) 1.71 (± 1.79)	139	9728	0.92
Sanity (Q11) 0.26 (± 1.85)	139	Membership (Q13) 1.71 (± 1.81)	139	5551	**0.0**
Sanity (Q11) 0.26 (± 1.85)	139	M. Stealing (Q15) 1.71 (± 1.79)	139	5420	**0.0**
Membership (Q13) 1.71 (± 1.81)	139	M. Stealing (Q15) 1.71 (± 1.79)	139	9616	0.95

B. *Statistical tests from Section V-C*

We first consider the results from Section V-C about our participants as in the order of the paragraphs in the main paper.

Role in team We now split our data according to whether a participant has a technical role (left side) or not (right side).

Non-technical		Technical			
μ (±sdt.)	#	μ (±sdt.)	#	U	p
Exposure (Q2) 0.34 (± 1.02)	91	0.81 (± 1.68)	48	1970	0.15
Poisoning Rating (Q5) 2.46 (± 1.26)	91	2.21 (± 1.51)	48	2370	0.39
Evasion Rating (Q7) 1.92 (± 1.7)	91	2.12 (± 1.41)	48	2135	0.83
Backdoor Rating (Q9) 1.79 (± 1.7)	91	1.69 (± 1.77)	48	2290	0.63
Sanity Rating (Q11) 0.03 (± 1.78)	91	0.69 (± 1.92)	48	1756	0.05
Membership Rating (Q13) 1.73 (± 1.74)	91	1.67 (± 1.93)	48	2151	0.89
M. Stealing Rating (Q15) 1.79 (± 1.78)	91	1.54 (± 1.79)	48	2340	0.48

Prior Knowledge in AML. We now divide our sample along the self-reported knowledge of AML (present, Sample I, not present: Sample II). For the security approaches(Q4), we encode the replies using the amount of implemented approaches (e.g., 0-7).

Sample I		Sample II			
μ (±sdt.)	#	μ (±sdt.)	#	U	p
Exposure (Q2) 0.52 (± 1.3)	40	0.49 (± 1.31)	99	1991	0.94
Security approaches (Q4) 3.45 (± 2.02)	40	2.72 (± 2.05)	99	2403	**0.05**
Estimated Risk (Q3) 5.5 (± 2.26)	40	4.93 (± 2.35)	99	2227	0.25
Poisoning Rating (Q5) 2.88 (± 0.95)	40	2.17 (± 1.44)	99	2533	**0.01**
Evasion Rating (Q7) 2.65 (± 1.22)	40	1.73 (± 1.67)	99	2580	**0.0**
Backdoor Rating (Q9) 2.38 (± 1.58)	40	1.51 (± 1.72)	99	2567	**0.01**
Sanity Rating (Q11) 0.52 (± 1.96)	40	0.15 (± 1.79)	99	2183	0.33
Membership Rating (Q13) 2.4 (± 1.51)	40	1.42 (± 1.84)	99	2581	**0.0**
M. Stealing Rating (Q15) 2.35 (± 1.28)	40	1.44 (± 1.89)	99	2460	**0.02**

ML knowledge and education. We now divide our sample along the self-reported knowledge of ML (present, Sample I, not present: Sample II). For the security approaches(Q4), we encode the replies using the amount of implemented approaches (e.g., 0-7).

Sample I		Sample II			
μ (±sdt.)	#	μ (±sdt.)	#	U	p
Poisoning Relevance (Q5)					
2.55 (± 1.08)	114	1.56 (± 2.04)	25	1738.0	0.07
Evasion Relevance (Q7)					
2.16 (± 1.49)	114	1.24 (± 1.9)	25	1811.0	**0.03**
Backdoor Relevance (Q9)					
1.88 (± 1.64)	114	1.2 (± 1.96)	25	1685.0	0.15
Sanity Relevance (Q11)					
0.27 (± 1.83)	114	0.2 (± 1.96)	25	1480.0	0.76
Membership Relevance (Q13)					
1.87 (± 1.74)	114	0.96 (± 1.93)	25	1811.5	**0.03**
Model Stealing Relevance (Q15)					
1.85 (± 1.71)	114	1.04 (± 1.97)	25	1769.5	0.05

To gain insights on the influence of education, we ran a regression model optimized with bfgs. The predictor was education (Q29), the predicted variable exposure (Q2). The obtained log-likelihood was -96.77, the AIC 205.5 and the BIC 223.1. The model had 133 residuals and 6 degrees of freedom.

	coef	std err	z	P> |z|	[0.025	0.975]
Education	0.133	0.096	1.38	0.17	-0.06	0.32
0.0/1.0	1.447	0.395	3.66	0.00	0.67	2.22
1.0/2.0	-1.5	0.364	-4.12	0.00	-2.22	-0.79
2.0/3.0	-1.373	0.393	-3.49	0.00	-2.14	-0.6
3.0/4.0	-2.956	0.991	-2.98	0.00	-4.9	-1.01
4.0/5.0	-2.147	0.694	-3.09	0.00	-3.51	-0.79

We used an analogous model to test the influence of education on estimated risk. The obtained log-likelihood was -210.71, the AIC 433.4 and the BIC 451. The model had 133 residuals and 6 degrees of freedom.

	coef	std err	z	P> |z|	[0.025	0.975]
Education	0.046	0.057	0.81	0.42	-0.07	0.16
-1.0/1.0	-2.023	0.342	-5.91	0.00	-2.69	-1.35
1.0/2.0	0.321	0.199	1.61	0.11	-0.07	0.71
2.0/3.0	-0.175	0.135	-1.29	0.2	-0.44	0.09
3.0/4.0	-0.086	0.132	-0.65	0.51	-0.34	0.17
4.0/5.0	-0.452	0.234	-1.93	0.05	-0.91	0.01

Gender. In this section, we divided the sample into female (left) and male (right) participants.

Female		Male			
μ (±sdt.)	#	μ (±sdt.)	#	U	p
Estimated Risk (Q3)					
2.7 (± 0.95)	20	2.46 (± 1.26)	99	1098	0.43
exposure (Q2)					
0.4 (± 1.16)	20	0.6 (± 1.43)	99	945	0.64
Poisoning Rating (Q5)					
2.1 (± 1.34)	20	2.48 (± 1.24)	99	846	0.28
Evasion Rating (Q7)					
1.65 (± 1.53)	20	2.11 (± 1.59)	99	811	0.19
Backdoor Rating (Q9)					
0.55 (± 1.63)	20	1.99 (± 1.62)	99	517	**0.0**
Sanity Rating (Q11)					
-0.15 (± 1.53)	20	0.31 (± 1.91)	99	867	0.37
Membership Rating (Q13)					
1.25 (± 1.76)	20	1.83 (± 1.79)	99	793	0.16
M. Stealing Rating (Q15)					
0.85 (± 1.96)	20	1.94 (± 1.72)	99	673	**0.02**