# "I Never Thought About Securing My Machine Learning Systems": A Study of Security and Privacy Awareness of Machine Learning Practitioners

FRANZISKA BOENISCH*, Fraunhofer AISEC, Germany
VERENA BATTIS*, Fraunhofer SIT, Germany
NIKOLAS BUCHMANN, Freie Universität Berlin, Germany
MAIJA POIKELA, Fraunhofer AISEC, Germany

Machine learning (ML) models have become increasingly important components of many software systems. Therefore, ensuring their privacy and security is a crucial task. Current research mainly focuses on the development of security and privacy methods. However, ML practitioners, as the individuals in charge of translating the theory into practical applications, have not yet received much attention. In this paper, the security and privacy awareness and practices of ML practitioners are studied through an online survey with the aim of (1) gaining insight into the current state of awareness, (2) identifying influencing factors, and (3) exploring the actual use of existing methods and tools. The results indicate a relatively low general privacy and security awareness among the ML practitioners surveyed. In addition, they are less familiar with ML privacy protection methods than with general security methods or ML-related ones. Moreover, awareness correlates with the years of working with ML but not with the level of academic education or the field of occupation. Finally, the practitioners in this study seem to experience uncertainties in implementing legal frameworks, such as the European General Data Protection Regulation, into their ML workflows.

CCS Concepts: • **Security and privacy → Usability in security and privacy**; Privacy protections; • **Computing methodologies** → *Machine learning*.

Additional Key Words and Phrases: Machine Learning, Security, Privacy, Awareness, GDPR

**ACM Reference Format:**
Franziska Boenisch, Verena Battis, Nikolas Buchmann, and Maija Poikela. 2021. "I Never Thought About Securing My Machine Learning Systems": A Study of Security and Privacy Awareness of Machine Learning Practitioners. In *Mensch und Computer 2021 (MuC '21), September 5–8, 2021, Ingolstadt, Germany*. ACM, New York, NY, USA, 39 pages. https://doi.org/10.1145/3473856.3473869

## 1 INTRODUCTION

Machine Learning (ML) is a rapidly growing area within the field of Artificial Intelligence (AI). Since 2012, the average rate of students being enrolled in university courses concerning this topic has more than tripled [9, p. 49]. Several countries in the world have put an AI strategy into place, *e.g.* Finland aims at training 1% of their population in the field [9]. Due to the increasing amount of personal data being collected, the importance of the topic continues to grow.

Since the early 2000s, several governments have put data protection regulations into place in order to protect the privacy of individuals whose data is being collected. The most prominent examples include the *Canadian Personal*

[*] Authors contributed equally to this research.

*Information Protection and Electronic Documents Act* (PIPEDA) [63], the *European General Data Protection Regulation* (GDPR) [91], and *California's Consumer Privacy Act* (CCPA) [80]. Those regulations all include rules for handling personal data. Yet, in general, as well as in ML-related workflows, functionality is still the most important factor when designing new products or models [59, 60]. *Privacy*, *i.e.*, protection against unwanted collection, use, and disclosure of personal data, or *security*, *i.e.*, ensuring the three IT security objectives of confidentiality, availability, and integrity [79], are, generally no significant driving factors [50]. They might instead be entirely neglected when deadlines approach [53]. As ML is increasingly used in security-critical applications, such as intrusion detection systems [18], and privacy-sensitive domains, such as medicine [82], this can have severe consequences. Moreover, for many years, research in the area of ML has mainly focused on finding new techniques to build increasingly powerful models. Only in recent years has the field of private and secure ML has seen an upsurge with the introduction of several techniques that allow for privacy preserving and secure analyses (see Section 2 for an overview of current research and developments).

Nevertheless, the final architecture and workflow of ML models—from the proof-of-concept to the final deployed product—are still determined by the human developers planning and implementing them. Since the threat space and vulnerabilities of the models against specific attacks depend on technical implementation details, the actual security and privacy rely, to a great extent, on the ML practitioners' *awareness* of potential risks and threats to their models. There exist various definitions of the term awareness in the context of information security [38]. They include the individuals' knowledge about security threats, countermeasures and precautions, and the understanding of the importance of the topic [38]. Without such awareness, the best methods are of little use if practitioners do not know about them, or consider them irrelevant, and hence do not implement them (correctly).

To the best of the authors' knowledge, no research has been conducted to investigate the individual security and privacy awareness of ML practitioners and their respective methods used. However, this is a crucial step on the way towards more secure and private ML, because such research not only indicates to what extent current tools and methods are successfully applied, but also unveils the challenges and improvement opportunities in the ML education. The present study contributes to this goal by answering the following two research questions (RQ):

1. *RQ1:* How is ML security and privacy awareness built, and which conditions contribute to the degree of knowledge with respect to threats and corresponding defenses?
2. *RQ2:* What is the current state of affairs concerning ML security and privacy among ML practitioners?

To answer *RQ1*, different aspects of awareness in secure and private ML are studied. This includes the practitioners' knowledge acquisition as well as the identification of factors contributing to awareness. To study *RQ2*, security and privacy attacks and practices are clustered according to how well they are known by and how widely they are used among the survey participants. Additionally, the ML practitioners' experience with selected libraries to support private and secure ML development is investigated. Finally, the influence of the introduction of juridical regulations on ML workflows is examined using the GDPR as an example.

The investigation of these questions shall serve as a basis to better support ML practitioners in privacy and security issues, to design helpful tools and standards, and to improve existing methods or develop new ones. Thereby, this study closes a critical gap on the way towards bringing the research on private and secure ML from theory to practice. The four main results of the study can be summarized as follows:

- The average awareness of security and privacy threats and protection measures among the surveyed ML practitioners is comparatively low.
- Academic education seems to have no significant impact on awareness of secure and private ML.

- ML protection methods put into place, especially for improving privacy, are less well-known than traditional and ML-specific security measures.
- The introduction of the GDPR appears to have no far-reaching impact on ML workflows in particular, and leaves the studied ML practitioners with several uncertainties.

This work is structured as follows. In Section 2 related work on security and privacy in ML and on studying practitioners' privacy and security practices in general is presented. Section 3 depicts methodological realisation, the preliminary pilot study, the structure of the questionnaire, participant acquisition, data analysis methods, and potential limitations and biases. The results of the study are shown in Section 4. Section 5 and 6 present the discussion, and conclusion with outlook, respectively.

## 2 BACKGROUND AND RELATED WORK

As context for this study, first, an overview of the current state of research in secure and private ML is provided. The threats, attacks and security practices described therein served as guidance for the design of the initial questionnaire. Second, relevant existing studies in the field of ML-related and general privacy and security practices from a developer's view are presented to put this work into context.

### 2.1 Secure and Private Machine Learning

ML has different security requirements than classic software [66]. Mainly, this is due to it being a rather new domain where the underlying security or privacy problems have not yet been fully explored or understood. This situation is aggravated by the fact that the well-known practices that support security integration in classic software development (*e.g.* static code analysis or code coverage) only partially address the issues in ML [66]. For instance, the prediction of an ML model does not only depend on the code used to train it, but also on its training data. While the pre-processing of data can be monitored by traditional security methods, other steps in ML applications use a whole new technology that creates new attacks, which in turn require different security mechanisms.

Research on making ML more secure and private focuses on the one hand, on the identification of risks and the development of concrete attacks against the models, *e.g.* data poisoning [14], model inversion [31, 88, 96], model evasion [13], and impersonation [3, 52]. On the other hand, it is concerned with the development of new methods, tools, and libraries for integrating security and privacy into ML [23, 32, 55, 93]. These methods can be roughly grouped by their purpose as follows:

- In order to improve data privacy in ML, a broad spectrum of defenses has been put into place, *e.g. Differential Privacy* [26], *Federated Learning* [16], *Privacy Preserving Record Linkage* [86], *Homomorphic Encryption* [15], and *Secure Multiparty Computation* [51], just to name a few. Most tools or libraries implementing them are rather new and exhibit a limited usability for non-experts.
- Similarly, the improvement of explainability and transparency is targeted, hence, making it easier to understand why a particular prediction is made [56].
- Further methods aim at making the models more robust against manipulation [40, 69].
- Besides the creation of methods specific to ML, a large part of research focuses on the adaptation of general security practices in order to be applicable to ML. These parts consist of, for example, *tracking the provenance* of the data (manually or automatically) [33], *restricting access* to the models [53], and *watermarking* the models in order to prevent model theft [84].

See [68] for a systematization of knowledge in the area of private and secure ML, and [6, 7] for further insights on the topic of security in ML.

### 2.2 Studying Developers' Privacy and Security Practices

Studying developers, *i.e.*, people who design, implement, or maintain program code, is a challenging task, and so far, not very common. Therefore, studies concerning developers are more often based on small sample sizes or rely on university computer science students [2, 50, 59, 60]. Current research suggests that students can indeed be used as proxies for professionals within such studies [12, 39, 74, 83]. Salman *et al.* [74], compared students and developers for (non-security-related) tasks and found, that, if students and developers are equally experienced in the topic, their code quality is comparable.

Studying individuals in a very specific field, such as private and secure ML, is even more challenging due to the small pool of potential participants. Very few papers exist with a similar focus as the study presented in this paper. The most closely related work to date was done by Kumar *et al.* [50]. To evaluate whether companies are equipped to protect and detect attacks on their ML systems, the authors conducted interviews with two employees each from a total of 28 organizations: the developer responsible for creating ML models and the security personnel responsible for securing the company's infrastructure. The study revealed that most companies do not possess tools or know-how to protect their ML systems and that most companies still focus on traditional security. Additionally, the authors found that attacks that can lead to a potential privacy breach are considered particularly dangerous by companies. Unlike the present study, Kumar *et al.* focus on companies and their general ML workflows rather than on the implementing developers and practitioners and their perception of the importance of this issue. Furthermore, they put more focus on emphasizing gaps in the security of the technical workflows instead of identifying factors that have led to these gaps—*i.e.*, awareness or rather the lack thereof among individuals. In their study, no real differentiation is made between security and privacy.

Within the broader field of general developer studies, there is also a small body of work related to studying developers and their security and privacy relevant coding practices.

Acar *et al.* [2] conducted an online study with Python developers in order to investigate the impact of using security APIs on code security. The developers had to fulfill specific coding tasks and answer a questionnaire. The results showed that while the self-reported status of a developer (professional or student) did not have a significant impact on the functionality and security of the code, the years of experience in development did. Each year of experience corresponded to an approximate 10% increased likelihood for producing functional code and a 5% increased likelihood for implementing secure results. Still, the level of security of code was quite low for both developers and students.

Naiakshina *et al.* [59] carried out a coding task study on 20 and a survey on 40 computer science students [60], respectively. Their main findings were that in the development process, in general, the participants considered functionality before security. The use of security measures increased significantly only when the participants were explicitly asked to implement them. Also, the knowledge of security practices did not guarantee a secure software implementation. The authors, furthermore, suggested that the use of good APIs is not sufficient if security works in an opt-in fashion, where the secure option is not default but needs to be set explicitly.

Further research in developer security and privacy practices focuses specifically on the development of apps.

Balebako *et al.* [5] surveyed around 200 app developers regarding their privacy practices and awareness. They found that most developers considered privacy policies hard to read, and criticized that they are mostly created without input from developers. Also, most of the developers indicated that they had not received any formal training in privacy and learned it on demand when being confronted with a new task that required it. In this case, they would rely on their social network and on specialists around them for input on how to apply it. The additional findings of their survey highlight that functionality is taken care of before privacy, and that developers' awareness on privacy is no indicator for the quality of privacy in the actual solution.

Nadi *et al.* [57] found in a developer study among 11 Java developers that simple APIs, and good documentation help developers to improve the security of their code. Jain and Lindqvist [41] showed in a lab study that by introducing appropriate APIs, developers could be nudged into using privacy-preserving programming choices.

Unlike the above-mentioned studies that are concerned with investigating general security and privacy practices among (app) *developers*, this study focuses on ML *practitioners*. For a precise definition of the term ML practitioner and the delimitation to the term developer, see Section 3.3.

## 3 METHOD

To investigate the privacy and security awareness and practices of ML practitioners, a study was conducted including questions about practitioners' demographics and working environments, and their knowledge of attacks and methods for implementing secure and private ML.

The study was conducted in the form of online questionnaires in two separate phases: a preliminary pilot and the actual study. The pilot served to evaluate the validity of the survey instrument and to inform the actual study of potentially missing aspects.

LimeSurvey [34] was used to host the online questionnaire. The participants did not receive any compensation for their participation.

Before being directed to the questionnaire, each participant was informed about the collection and handling of questionnaire data. The participants were also informed that their participation was voluntary and that they could discontinue their participation at any point in time. Only participants who gave their consent were forwarded to the questionnaire.

### 3.1 Pilot Study

As already mentioned, studying a niche populations such as ML practitioners, is no easy task. This is because of the relatively small population size and the rapid advancements in the research field of ML in general. Due to this, a preliminary pilot study was conducted in order to get an as accurate as possible impression of the current state of matters. Participant acquisition was performed through sending a link to the questionnaire via email to contact persons at AI-related companies, resulting in 41 completed questionnaires from eleven different European countries. The questionnaire was furthermore given out to 40 students enrolled in a local university's ML course. Out of the 40 students, 32 completed the questionnaire.

With the pilot study, an evidence-based-design approach (EBD) was followed, *i.e.*, instead of relying solely on information previously published in academical papers, heavy use of qualitative elements in the sense of free-text fields was made.

The free-text questions were used to gain qualitative insights into the practitioners' experiences and thoughts about security and privacy in ML and the introduction of the GDPR. Those aspects would not necessarily be captured through closed questions in the questionnaire. Also, biased answers, *e.g.* that some participants might know the attack or defence itself but not the correct associated name, should be avoided. Hence, the participants were asked to describe, in their own words, what threats to ML models they were aware of and what kind of measures they know to counter the aforementioned threats. The selection of the ML libraries depicted in Figure 10 followed the same procedure.

Other free-text fields served to collect further input from the participants, *e.g.* whether answer options were incomprehensible or even missing. For the evaluation of all free-text fields, code sets for the different question groups were identified by two researchers separately. After agreement on a final code set, the codes were applied independently by the two researchers to the data of the pilot study. To evaluate the rate of agreement between the two raters, Cohen's Kappa for inter-rater reliability was calculated [21]. The obtained Kappa value of 0.93 indicates that the raters had a strong agreement on the codes, and the labels represent the data well [30]. However, in order to draw conclusions on the data, the disagreements on the categorization were resolved by discussing each case independently until a congruence was reached.

Additional quantitative study elements, such as multiple choice questions, were used to find patterns in ML practitioners' security and privacy practices, and to identify relationships between their demographics, awareness, and privacy and security-related behavior.

### 3.2 Structure of the Final Questionnaire

Based on the findings from the pilot study, the questionnaire was adapted for the main study. It consists of six main groups of questions, and is attached at the end of the Appendix Section.

(1) Demographics: This section captures the demographic background of the participants, *i.e.*, their education, the country they were working in, their daily ML-related tasks, and their present working situation. Participants who indicated that they were currently employed were asked additional questions about their employing company *e.g.* number of employees, how ML is applied, and the sector in which the company operates.
(2) Data and Sensitivity: The practitioners were asked, among other things, what type of data they were dealing with, whether this data is (directly) related to individuals, as well as which domain it stems from.
(3) ML security: The questions cover how important the participants judged securing their ML models, how they built their knowledge on ML privacy and security, and who in their working environment is responsible for securing the ML models.
(4) Attacks: In this question group, four attacks (inversion [31], impersonation [3, 52], poisoning[14] and evasion attack[13]) on the privacy and security of ML models were presented. For each of the attacks, the participants were asked to indicate whether they were familiar with this attack, and whether they had already implemented preventive measures to defend against the respective attack. To avoid participants mistakenly marking an attack as unknown just because they were unfamiliar with the particular keyword, a short explanation of the attack was provided together with its name.
(5) ML privacy and security practices: Within this section, first, eight ML privacy and security libraries were presented and the participants were asked whether they were familiar with these libraries, and whether they had used them. Furthermore, 14 security and privacy practices identified within the participants' answers in the pilot study were presented, along with an explanation for each of them. Again, the practitioners were asked to specify per method, whether they were familiar with it, and whether they had already implemented it.
(6) GDPR: The last section contained questions regarding participants' familiarity with the GDPR, and the changes in their ML-related privacy practices caused by its adoption.

The questionnaire consisted of 25 questions and took 11 min. 18 sec., on average, to complete. Both questionnaires, the preliminary pilot study as well as the main study, were tested for the validity of the survey instrument before they were applied in the field. This testing was done separately and independently of each other and on the basis of a group of six subjects.

### 3.3 Participants

The actual study took place from July 2020 until October 2020. To recruit participants, the questionnaire was promoted through the official channels of the authors' institutions, such as websites, and social media accounts (LinkedIn, Twitter, and Facebook). In order to reach as many international participants as possible, the link—together with a short description of the survey—was posted in various ML-related groups on Reddit and LinkedIn. In addition, the link for the questionnaire was shared by the authors through their own LinkedIn profiles as well. Between July and October, 1471 volunteers clicked on the link and were directed to the survey, of which 94 completed the questionnaire in full. As the aim of this study was to investigate practitioners' awareness, all students were removed from the dataset, reducing the number of completed questionnaires to a total of 83. For participant details, see Table 2 in the Appendix.

The majority of participants held high educational degrees, with 80 (96%) of them having at least a bachelor's degree. The level of education is consistent with what was previously found among data scientists in a survey by Kaggle [44] (91% ).

73 (88%) of the participants were employed and 49 (59%) in the early stages of their ML career—considering their work experience with ML. This is also consistent with Kaggle [44] where 55% of the participants have less than three years experience. This survey reached more participants working in larger companies (54, 65% work in companies of over 200 employees) than medium-sized (18, 22%) companies with 11-200 employees, or smaller ones (5, 6%) with ten or fewer employees.

The participants' working domains were most often related to *customers and users* or *smart environments and connected devices* (for a summary of the environments ML is applied in, see Table 3 in the Appendix). The most frequent types of data handled by the participants were *images*, *sensor*, *tabular*, or *text data*. Within the 'other' option, several participants specified working with *industrial and manufacturing data*, *publicly available datasets*, or *education-related data*.

Additionally, more than half of the participants stated that ML is the main component of the products developed by their department (47, 56.7%), whereas one-third declared that ML is included in the products developed but not as a key element (28, 33.7%).

A total of 59 (71%) participants, and thus the majority, stated that their daily tasks included applying ML libraries, such as TensorFlow [1] and Scikit-learn [70], or conducting data analyses (54, 65%). Roughly half of the participants indicated, that their tasks included evaluation (47, 56.6%), data cleansing and preparation (45, 54%), coordinating ML projects and workflows (44, 53%), as well as developing custom ML applications, *e.g.* designing custom neural networks for given tasks (36, 43.4%). These three questions regarding daily tasks, data domain and data type were posed as multiple choice questions, where the participants could select all applicable options.

The studies mentioned in the related work part (Section 2.2) examined the awareness of *developers* and not just *practitioners*. This study addressed ML *practitioners* and asked them to specify their daily ML-related tasks. Based on the responses, the group 'ML developers' was artificially created ex post. The authors consider participants to be ML developers if they stated that they either 'develop custom ML applications (*e.g.* designing custom neural networks for a given task)' or that they 'develop ML tools or libraries from scratch'. This ex post definition counteracts the fact that, since there is no clear definition what an ML *developer* constitutes of, some people, if asked in a questionnaire whether they consider themselves an ML developer, would deny it.

### 3.4 Data Analysis Methods

The data export from the LimeSurvey online questionnaire and the data preprocessing was implemented in Python [85]. For the analysis, the Python-libraries *scipy.stats* [90] and *factor_analyzer* [76] were used. The notebooks written for pre-processing, analysis and visualization of the results are can be obtained from the authors.

The analysis itself was divided into two parts. The first part focused on identifying correlations and dependencies within the data. Correlations were assessed using *Spearman's rank correlation coefficient*, while a $\chi^2$ *test of independence* was used to determine whether two categorical variables exhibit a significant relationship. In the second part of the analysis, group comparisons were conducted to investigate whether differences in the response variables can be explained by being affiliated to a particular group. Comparisons of three or more groups were performed using the *Kruskal-Wallis H test* [49]. If the Kruskal-Wallis test indicated a significant difference between the distributions of the tested groups, or if the analysis of interest was to compare just two groups in general, a pairwise comparison employing the *Mann-Whitney U-test* [28] was performed. Since all tests were performed on a single data set, the problem of multiple comparisons was addressed by applying the *Benjamini and Hochberg correction* [11], which is performed within individual hypothesis families. In this paper, the hypothesis families were formed based on the research question that each hypothesis attempts to answer. Therefore, all hypotheses in Table 4 (see Appendix) form one family, and all hypotheses in Table 5 (see Appendix) form the second. Moreover, in the present work, the *corrected p–value*, $p^*$, is reported together with the respective test statistic rather than the original, uncorrected p–value, $p$.

One of the key questions of this work is how awareness is built up among ML practitioners regarding the privacy and security of ML and which aspects play a role in this process. To answer these questions an *exploratory factor analysis* was carried out in order to estimate the latent construct *awareness*. In the preliminary pilot study, participants were asked to use their own words to describe threats to ML and possible defences. This procedure resulted in the identification of four attacks and 14 defenses. Following [38], stating that awareness constitutes of the individuals' knowledge about security threats, countermeasures and precautions, these 18 items were expected to correlate strong enough with *awareness* to be used as proxies in estimating it (see Appendix Table 1). To verify the adequacy of employing a factor analysis, *Bartlett's test* of sphericity [8] ($\chi^2_{(2)} = 383.91$, $p < 0.001$) was conducted and supplemented by the *Kaiser-Meyer-Olkin test* [45, 46] ($KMO = 0.84$) to measure the suitability of factor analysis. Both criteria indicate that it is appropriate to use a factor analysis on this subset. Hereafter, the number of factors to be selected was determined on the basis of a scree-plot, which indicated that a one–factor solution was sufficient, explaining 34% of the variance. Since factors were expected to be uncorrelated, a varimax rotation was performed. Only items with loadings $> 0.45$ were considered to correlate strongly enough with the factor. As a consequence, five items were excluded from further analysis (see Appendix Table 1). To test the reliability of the constructed scale, *Cronbach's Alpha* [22] was computed. The internal consistency of the scale, which now consists of 13 items, is 'good' by the statistic's definition ($\alpha = 0.86$).

Factor scores were calculated to derive a single variable, which truthfully quantifies the survey participants' level of awareness. These factor scores represent estimates of the unobservable characteristic *awareness*, hence, *the hypothetical value each participant would exhibit if awareness was directly measurable*. The estimated scores ranged from $-1.55$ to $2.21$. To facilitate interpretation, the scores were normalized to the range $[0, 1]$. From here on, they are referred to as *Awareness Scores*.

The endpoints of this continuous scale, 0 (no awareness) to 1 (high awareness), are no absolute numbers but rather relative ones, since the estimated factor scores of the survey respondents were used for normalization. Thus, an awareness score of 1 does not mean that the respondent knows everything about private and secure ML, but rather that this respondent was the person with the highest awareness of privacy and security threats and solutions in the context of ML in this survey. The same holds true in an opposite manner for the minimum point of the normalized scale, 0.

### 3.5 Limitations and Biases

Studying ML practitioners is a challenging task because they are a specialized population that is difficult to recruit. Therefore, this study may have some limitations that need to be considered before interpreting the results.

The sample contains some demographic biases; the first being sampling bias. Although great effort was taken to recruit an international sample, the majority of the participants indicated, that they are currently working in Europe. Additionally the sample is not balanced with regard to the fields where the participants apply ML (see Figure 2 on page 7 and Table 2 in the Appendix). In terms of years of professional experience with ML, the group of participants who state that they have 1-3 years of work experience in the respective field is highly over-represented with 49 (59%) answers. As mentioned earlier, this observation mirrors the results of the Kaggle survey [44] (55%). This is not necessarily a bias, but could rather reflect the current dynamics of ML, which leads to many new positions being created continuously. Yet, the results may not be representative of all ML practitioners in all fields and at all levels of experience.

Not offering monetary compensation for participation further contributed to the difficulties in the recruitment. The decision to not offer monetary rewards was deliberate on the part of the authors, as they did not want to attract dishonest participants who just click through the study, to collect their reward at the end. The practitioners' decision to participate in the study, therefore, depended solely on their individual motivation and interest in the topic. Weighing the pros and cons, the authors accepted the risk of a potential bias of the sample towards an over-representation of developers who are already more interested in and aware of the subject.

The authors are aware that conclusions about the characteristics of the population, based on this survey, are restricted due to the above limitations. However, they would like to point out that the results of the conducted study do not intend to claim general validity, but rather to serve as a starting point to determine the current state of awareness among practitioners, as well as to help them with better tools and standards. In particular, the qualitative results are independent of sample size and provide valuable insights into the field.

## 4 RESULTS

In the following, the two research questions presented in Section 1 are answered in a structured manner. As a general finding, it was observed that the awareness of ML privacy and security among the participants was relatively low. See Figure 1 for a representation of the distribution of awareness scores in the sample. The quantiles of the normalized awareness scores are $q_{0.25} = .173$, $q_{0.5} = .389$, $q_{0.75} = .522$, ranging from 0 (no awareness) to 1 (high awareness). When asked how they built their current knowledge of ML security, several participants in both the preliminary pilot and the actual study made use of the provided comment boxes to add supplemental information, such as *"I didn't yet work with sensitive data, so I didn't yet have to build that knowledge"*, or *"I never thought about securing my machine learning systems, mainly because they are research oriented rather than production oriented systems"*. These statements correspond with the finding that 24% of all participants had never heard of any of the four mentioned attacks (inversion, impersonation, poisoning and evasion attack), which aligns with the findings of Kumar *et al.* [50]. Moreover, securing ML models

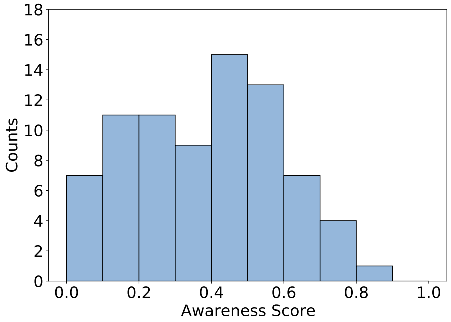


Fig. 1. Count plot of the participants' awareness score. The values were normalised to the range from 0 (no awareness) to 1 (high awareness).

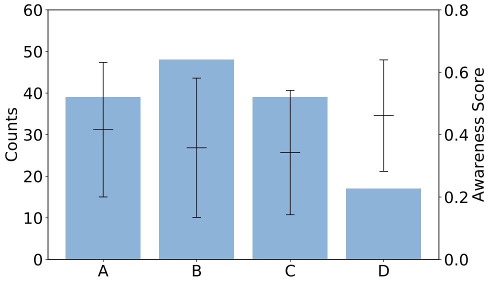


Fig. 2. Absolute number of participants per occupational field in connection with the average awareness value and standard deviation resulting in the respective field. A: 'Industry', B: 'Academic Research', C: 'Industrial Research', D: 'Hobby'. Please note that the bars do not add up to the sample size of 83 due to the possibility of multiple answers.

against these threats does not seem to be a driving factor in most of the participants' working environments either (see Section 4.2).

### 4.1 Elements influencing Awareness

The goals and workflows in industry are demonstrably different from those in academia. Academics typically strive to outperform a specific benchmark, usually using fixed training data sets. Professionals in industry, in contrast, have a fixed performance target—what data and model is used is secondary, as long as the target is met. One would expect that this contrast also reflects in the level of individual awareness motivating the working hypothesis that the field of occupation does play a role in the extent to which individuals are aware of certain threats to their ML models. In addition to industry and academic research, the fields of industrial research and hobby are also considered. Due to the possibility of specifying several fields of occupation, neither group-wise comparisons nor Friedman tests are applicable here. Figure 2 depicts the absolute numbers of participants who assigned themselves to a specific occupational field. Although working in industry seems to result in a slightly higher average awareness compared to working in research (academic or industrial), no considerable differences in the distribution of the average awareness scores can be observed.

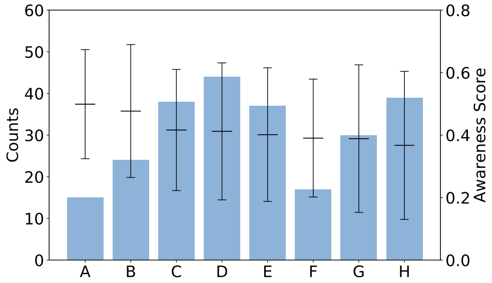


Fig. 3. Absolute number of participants per data type in connection with the average awareness value and standard deviation resulting per respective type. A: 'Audio', B: 'Location', C: 'Tabular', D: 'Images', E: 'Text', F: 'Video', G: 'Metadata', H: 'Sensor'.

Since it must be assumed that spillover effects exist for people who work in several areas, this convergence of the mean values is not surprising.

In a similar vein, the question of whether ML practitioners' awareness differs depending on the data type used was explored. For example, it is reasonable to assume that someone working with medical image data or portrait images would be more aware of potential privacy threats than someone working with industrial sensor data. A visual inspection of the eight groups (image, video, audio, text, location, meta, sensor and tabular data) yielded only slight differences in their average awareness (see Figure 3). However, it is noteworthy that some groups, such as audio (.50) and location data (.48), have even higher mean awareness scores than image data which has an average awareness score of .41. One possible explanation could be that both audio and location data can be seen as particularly sensitive and thus more worthy of protection, which in turn has a positive influence on awareness. However, a similar argument can be made with respect to video or metadata, which, at least in this sample, had, apart from sensor data, the lowest mean awareness scores (.39 and .39). In the end, all data types, *e.g.* tabular, audio or visual, can contain both sensitive and non-sensitive information. Moreover, it can be assumed that practitioners come into contact with more than one data domain, making a clear attribution of an impact to one specific data type not possible.

A comparable analysis was conducted on the domain the data stems from. The result can be found in Appendix, Figure 11.

The finding that the participants exhibit a relatively low level of privacy and security awareness in the context of ML is quite surprising given the fact that 54 participants (65%) stated that ensuring the security of ML models is 'important' or even 'very important', hence backing up the findings of Naiakshina *et al.* [59, 60]. Personal perception of the importance to secure ones ML models might be a valid reason to seek education on the field. Yet, no correlation between the ML practitioners' perception of the importance and their awareness can be observed ($r_{(81)} = .073$, $p^* = .62$). Given the fact that the introduction text to the survey stated that the questionnaire aimed at researching security and privacy awareness, the participants' answers to the question of importance might also contain a desirability bias.

When considering education, given the strong increase in study programs teaching AI and ML [10], one might expect that those who have a higher educational degree find themselves in the right tail of the distribution depicted in Figure 1. However, the level of education seems to give no suitable indication of the level of an ML practitioner's awareness ($\chi^2_{(4)} = 0.89$, $p^* = .827$). This could be because universities have only started to offer these courses in recent years, so the impact is not yet directly tangible in the work environment. In fact, the amount of time participants

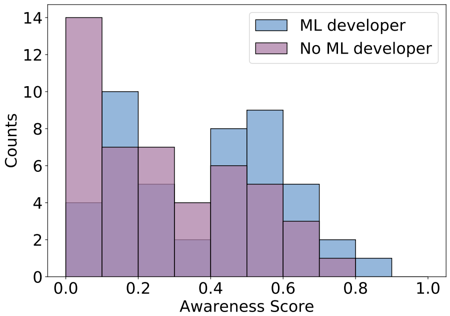


Fig. 4. Distribution of the ML practitioners' awareness score grouped by whether they are considered to be ML developers or not.

already work with ML, either professionally or as a hobby, has the strongest effect on the level of individual awareness ($r_{(81)}$ = .36, $p^*$ = .005).

Following Pieczu *et al.* [71], the working hypothesis was built, that the size of a company has an impact on the security awareness and practices of individuals, for example through the implemented security policies. Surprisingly, the size of a company does not seem to have an effect on the individual's level of awareness. When testing for differences in the distribution of awareness grouped by company size, no support for the working hypothesis could be found ($\chi^2_{(8)}$ = 15.26, $p^*$ = .109).

With hindsight to the privacy aspect of ML, the authors suspected that practitioners who are used to work with sensitive data (*e.g.* health data) might show a higher degree of awareness than those who work with data not related to individuals at all (*e.g.* weather data). One might expect the highest mean awareness among ML practitioners working with data that is directly related to individuals. However, practitioners working with data only indirectly related to individuals have the highest mean awareness (.44), compared to those working with data that is directly related to individuals (.35), or not related to individuals at all (.32). An additionally conducted Kruskal-Wallis test could not detect any significant differences in the central tendencies of the three groups ($\chi^2_{(2)}$ = 4.09, $p^*$ = .194), thus indicating that even working with data directly related with individuals has no effect on whether or not a practitioner is aware of certain threats.

In addition to the data type used, the authors postulate that the data domain and its sensitivity as well as the ML practitioner's daily ML-related tasks have an effect on awareness. In this context, it is crucial to distinguish between ML practitioners who merely apply libraries and those who actually develop them. As described in Section 3.3, the group 'ML developers' was created artificially by the authors based on the participants' day-to-day tasks at work. It was investigated whether developers can be expected to have a deeper understanding of machine learning and its intricacies and thus a higher awareness of the threat landscape than, for example, users who simply apply ML libraries to a specific task. Not surprisingly, the corresponding null hypothesis that there is no difference between both groups, was rejected. The group of ML developers did have a significantly higher awareness score than non-developers ($U_{(41,42)}$ = 584.0, $p^*$ = .018). For visualization see Figure 4.

When asked by what means they built up their current knowledge of ML security and privacy, only 25 (30%) respondents said they gained their knowledge at university. Most participants reported to have built their awareness through 'practice' (61, 73.5%) and 'self-study' (56, 67.5%) which is congruent with the findings of Balebako *et al.* [5].

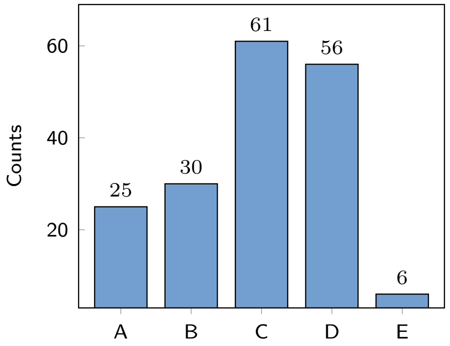


Fig. 5. Overview of the educational methods used by the ML practitioners to build knowledge about ML security and privacy. A: 'University', B: 'Workshops', C: 'Practice', D: 'Self-Study', E: 'Other'.

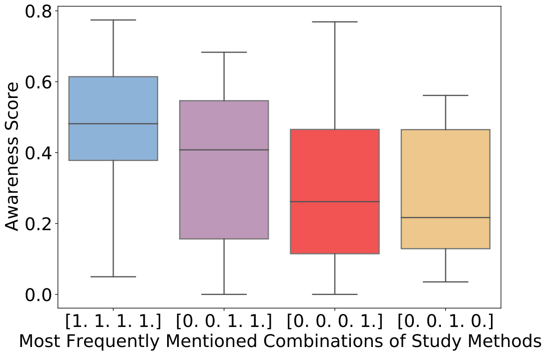


Fig. 6. Distribution of ML practitioners' awareness scores for the four most frequently mentioned combinations of learning methods used by the participants. The binary arrays decode for the following methods whether they were applied (1) or not (0): [University, Workshops, Practice, Self-Study]. The box plots correspond, from left to right, to 15, 14, 15 and 13 mentions respectively.

See Figure 5 for a consolidated overview on the participants' answers regarding the study methods used. Please note that in this question the participants were asked to check all answers that applied to them, hence the counts do not sum up to $n = 83$. The evaluation of the participants' answers is consistent with the general assumption that the more sources an individual uses for learning, the higher their level of awareness. Due to the possibility of specifying several learning methods used, it was not possible to determine the individual influence of a particular method on the level of awareness, *i.e.* whether a particular method contributes to practitioners becoming more aware of a particular attack. See Figure 6 for a more detailed depiction of the four most frequently reported combinations of learning methods and the respective awareness scores.

## 4.2 Working Environment and the GDPR

As the working environment specifies many working methods and general conditions for its employees, it might also have an impact on their security and privacy practices. When testing whether the company size and who is responsible for ensuring security of ML models are stochastically independent, the null hypothesis could not be rejected

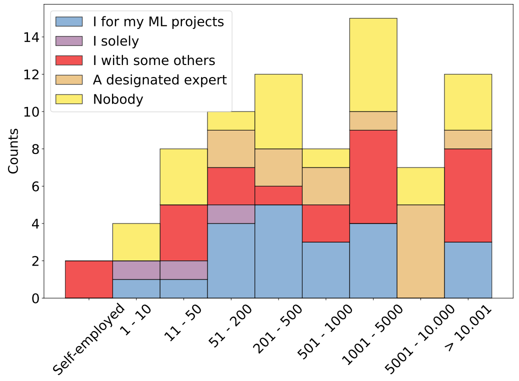


Fig. 7. Overview of the distribution of responsibility for securing ML models (*Who takes care of the security of the machine learning (ML) models in your working environment?*) by company size.

($\chi^2_{(32)}$ = 43.11, $p^*$ = .182). This result can also be seen in Figure 7. The plot indicates that irrespective of the size of a company, in the majority of the cases, ML security is taken care of, whether by an individual, a collective or a designated expert. However, the figure also depicts that there seem to be cases where nobody in the company is responsible for ML security, irrespective of its size. Some comments additionally given to this question included, that either the security is ensured through the processes that the participant's company has set up, or that whoever is in charge of model deployment, is also responsible for securing said models. Several participants also stated that their IT department, and in some cases security or data science teams are in charge of the security.

Also, no support was found for the working hypotheses of whether who is responsible for implementing security solutions depends on educational degree ($U_{(47,35)}$ = 774.5, $p^*$= .395) or the years a person is already working in ML ($U_{(48,35)}$ = 728.5, $p^*$ = .212). This may seem counter-intuitive since, as reported above, there is a positive correlation between the years practitioners work in an ML-related position and their awareness. One might therefore expect those with longer experience to be in charge.

Another factor that might influence the participants' ML security and privacy practices, are legal obligations.

This paper addresses this question using the GDPR as an example. As the GDPR may not apply to participants working outside EU member states, these participants were excluded for the following analyses. Therefore, the following figures only refer to the 64 ML practitioners who reported working in an EU member state.

First, the changes in the participants' ML-related workflows resulting from the introduction of the GDPR were examined. Figure 8 shows the survey participants' self-assessment of this very question, grouped according to whether the data they work with is directly related to individuals, indirectly or not related at all. A simple visual inspection of the graph shows that the majority of participants indicated that there were no changes at all. Unsurprisingly, within this group, the majority of practitioners works with non-personal data. However, no evidence of the opposite effect could be observed: the group working with personal data did not report great changes in their work processes either. Instead, this group and the one processing only indirect personal data behaved quite similarly. The authors suspect that this effect is observed because, according to the GDPR, also indirectly person-related data needs to be protected since there is a possibility for re-identification. The findings depicted in the graph are further supported by a Kruskal-Wallis test ($\chi^2_{(2)}$ = 12.39, $p^*$ = .005) accompanied by pairwise comparisons between the different degrees of data-sensitivity.

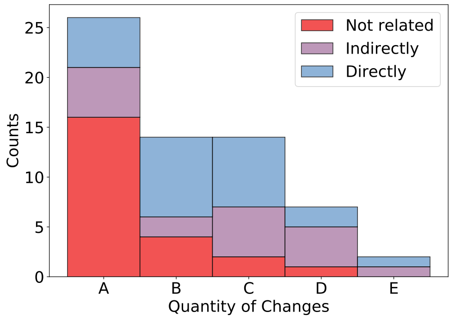


Fig. 8. Overview on the extent to which the introduction of the GDPR has provoked changes in the participants' ML workflows grouped by whether the data they are working with is directly related to individuals, indirectly or not related at all. A: 'Not at all', B: 'Very little', C: 'Somewhat', D: 'Very much', E: 'To a great extent'.

The distribution of practitioners working with non-human related data differs significantly from the remaining two ($U_{(23,17)} = 97.0, p^* = .005$), ($U_{(23,23)} = 130.0, p^* = .003$), while this is not the case when looking at the group working with indirectly and the one working with directly human related data ($U_{(17,23)} = 171.5, p^* = .38$).

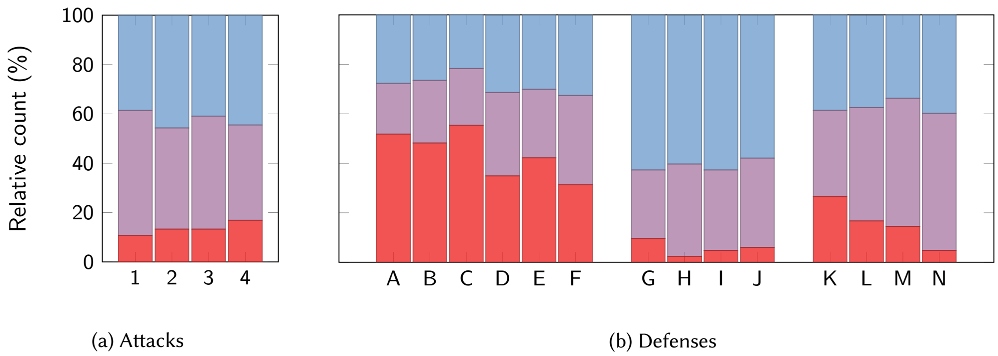
In order to gain deeper insights into the changes due to the GDPR and to explore the discomforts and uncertainties ML practitioners face as a result, several free-text questions were included in the survey. The participants' answers revealed that changes mainly affect general, and not specifically ML-related workflows. These changes include safe data storage (location and security measures of the servers, access control, encryption etc.), anonymization, more economic data collection, and more documentation. One participant, for example stated *"It [the GDPR] is more important when it comes to data collection than the training phase or ML model development"*. Only very few answers indicated changes in the ML workflows, for example one participant said *"I can not [sic] use a lot of features that I used before and so, the strategy to solve some problems has changed"*.

Yet, several ML practitioners expressed their insecurity concerning certain aspects of the GDPR in their ML workflows. Within these, insecurity about the implementation of the *Right to be Forgotten* were listed most frequently. Many ML practitioners seem to lack knowledge of adequate methods to implement this requirement, as the quote of one participant illustrates: *"Since GDPR requires the "right to be forgotten", we would need to artificially keep a correlation in order to be able to forget a specific individual, which, however, violates the GDPRs requirement to only store information that is required to operate a system"*.

A last finding from the GDPR-related questions highlights the importance and responsibilities of third-party infrastructure and service providers for privacy and security. For example, several participants stated relying on those third-parties' services without questioning them. One participant, when being asked about changes in the ML workflow caused by the GDPR, just indicated *"checked some boxes in AWS"*, another one explained more thoroughly *"I heavily rely on third part services to train and store my data, such as google cloud platform and AWS. Thereby, I hope their default security support is enough to protect my models and data"*.

### 4.3 Security Practices

The visual and statistical analysis of the ML practitioners' familiarity with security practices and methods for ML showed that there is a significant difference in acquaintance and experience with them (unfamiliar, familiar, has implemented)



(a) Attacks

(b) Defenses

Fig. 9. Overview of the participants' awareness among the methods presented in the questionnaire. For *attacks*, the question was if the participants had already implemented measures to defend against it, or if they were familiar, or unfamiliar with it. 1: Inversion Attacks, 2: Impersonation Attacks, 3: Poisoning Attacks, 4: Evasion Attacks. For *defenses*, the participants were asked if they had implemented this defense, or if they were familiar, or unfamiliar with it. Cluster 1 (A: Data Sanitization [65], B: Access Control [75], C: System Security [79], D: Ensemble Learning [24], E: Data Provenance [19], F: Observing Model Input at Inference Time [7]), Cluster 2 (G: Differential Privacy [27], H: Homomorphic Encryption [37], I: Watermarking [58], J: Privacy Preserving Record Linkage [87]), Cluster 3 (K: Smoothing Prediction Output [35], L: Introducing Delay [77], M: Adversarial Training [36], N: Federated Learning [48]).

($\chi^2_{(13)}$ = 93.19, $p^* < .001$). The participants' experience with the respective methods is depicted in Figure 9b. In order to obtain the clustering, the K-Means algorithm [42] was used. The number of clusters was determined by the elbow method [47], resulting in three semantic units.

The first cluster consists of methods that up to half of the developers indicated having already implemented. This cluster includes *data sanitization* [65], *access control* [75], *system security* [79], *ensemble learning* [24], *data provenance* [19], and *observing model input at inference time* [7]. All these methods can be considered classic security or standard ML methods, which would explain their prevalence.

The second cluster consists of methods that are typically used for privacy or intellectual property protection in ML models, such as *differential privacy* [27], *homomorphic encryption* [37], *watermarking* [58], and *privacy preserving record linkage* [87]. The majority of practitioners indicated not being familiar with these methods. Only some knew about them and very few have already implemented them.

The last cluster consists of specific ML methods that can be applied to secure ML models, such as *smoothing prediction output* [35], *introducing a delay for model interaction* [77], and *adversarial training* [36], or to protect training data, as in *federated learning* [48]. Similar to the second cluster, only few participants indicated that they have implemented the methods, however, the percentage of participants who stated that they were theoretically familiar with the methods, is higher.

In contrast to the protection measures, the familiarity of the participants with the four attacks did not exhibit significant differences ($\chi^2_{(3)}$ = 1.12, $p^*$ = .772). In general, more than half of the participants had at least heard of this type of attacks. Ultimately, however, only 10-20% of the respondents have already implemented an attack (see Figure 9a), *e.g.* to test whether the defence they used was strong enough to withstand an actual attack.

The survey participants were also asked about their familiarity with selected libraries for ML privacy and security. As mentioned above, the libraries were selected according to the answers given during the pilot study. The authors acknowledge that all of them are Python libraries, which they believe is due to the fact that Python is a very popular language for ML applications [92]. In addition, Kumar *et al.* [50] reported that 16 out of the 28 organisations they

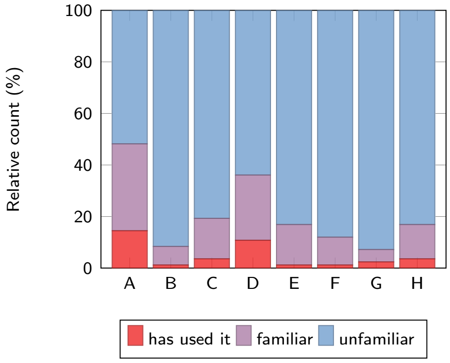


Fig. 10. Libraries to support privacy and security in workflows: A: TensorFlow Privacy [54], B: Cleverhans [67], C: PySyft [73], D: Googles Differential-Privacy [94], E: Uber SQl Differential-Privacy [43], F: AdverTorch [25], G: Foolbox [72], H: ART Toolbox. [61]

surveyed use Python frameworks such as Keras, TensorFlow or PyTorch. While ten rely heavily on ML-as-a-service providers and only two built their ML systems from scratch, relying neither on toolkits nor platforms. Figure 10 depicts the distribution of the participants' answers. It is apparent that across all libraries, the percentage of ML practitioners who have actually used them, is remarkably low. Only the tensorflow privacy library [54] achieves a publicity of around 50% among all participants.

## 5 DISCUSSION AND IMPLICATIONS OF THE FINDINGS

In this section, the results and their implications are discussed with regard to the study's two research questions.

RQ1: *How is ML security and privacy awareness built, and which conditions contribute to the degree of knowledge with respect to threats and corresponding defenses?* Studying ML security and privacy awareness and knowledge yielded three main results.

**The surveyed ML practitioners' awareness of threats, as well as of ML security and privacy practices is relatively low.** This finding aligns with the findings of Kumar *et al.* [50] and the general studies on security practices presented in Section 2.2. One possible explanation for this could be that in ML, similar to standard IT applications, security requirements tend to be kept separate from other system requirements [29] and that functionality is addressed before security [59]. One could also argue that most companies are currently in the early stages of adopting ML. Thereby, products are still more experimental and rather proofs of concept—with security and data protection playing a subordinate role. [53].

**Most ML practitioners had no academic training on ML security and privacy.** As depicted in Figure 6, only around one third of the participants obtained their knowledge from university. A similar tendency was already reported by Balebako *et al.* [5] for general security practices. With ML security and privacy being even more recent topics than general security or privacy, this finding is to be expected. Given the massive increase of student enrollments in ML-related courses within the last years, the overall literacy on the topic might increase in the future. However, so far, the results from this study suggest that the educational degree of the ML practitioners does not have a significant impact on their awareness in these topics, whereas years of work experience in ML do. A similar finding was presented by Acar *et al.* [2]. Therefore, extending academic education on the topics might be just one of various steps towards a higher awareness. This intuition is supported by the fact that the present study showed a positive correlation between the degree of ML security and privacy awareness and other sources of education and practice among the surveyed ML practitioners.

**The study showed no correlation between the size of the company the surveyed ML practitioners work in and their personal awareness.** The generalizability of this result, is, however, questionable. It might result from the demographics of the survey participants or the small sample size. In general, one would expect the awareness on security and privacy in larger companies to be higher, given that such companies often have company-wide coding and security guidelines, internal review processes, and the means to afford regular training for their employees. However, smaller companies or start-ups may be savvier about security and privacy threats because they have to deal with the problems themselves, whereas larger companies have designated security departments taking care of such issues.

RQ2: *What is the current state of affairs concerning ML security and privacy among ML practitioners?* Concerning the current state of affairs among ML practitioners, four main findings can be reported.

**The participants' familiarity with protection strategies for their ML models is very unevenly distributed.** Interestingly, the clusters of familiarity over the protection methods determined by the K-Means algorithm correspond well to their semantic units. The cluster of general security practices seems to be more broadly known and applied than the cluster of (partly) ML-related security measures. The cluster with the lowest familiarity contains specific methods for ML privacy protection. This indicates that these methods have, so far, received the least attention. This might reflect the general state in software engineering, where security development life-cycles are more established, whereas privacy engineering is just emerging. This leaves open the question of whether less attention is paid to ML privacy than to ML security, or whether privacy has so far been considered mainly in other phases of the workflow — as suggested by some participants' answers to the free-text question.

**The surveyed practitioners' familiarity with the presented security and privacy libraries for ML is low.** Another important aspect for the integration of security and privacy in ML workflows is the ML practitioners' use of dedicated tools or libraries. According to Wurster and van Oorschot [95], security tools should be more usable than insecure ones because it is impossible to forbid insecure solutions, however improved usability and understandable user-interfaces could encourage developers into choosing the more secure option [41, 81, 95]. The fact that, to date, many of the existing security and privacy libraries are not particularly usable, apply solely to very specific scenarios, and lack expressive documentation might be an explanatory factor for their low popularity. Therefore, in order to expand the use of ML privacy and security libraries in the future, it might be valuable to continue investigations on ML practitioners' current practices beyond the findings of this work and to study their functional needs thoroughly [78]. Additionally, human factors should also be taken into account in the evaluation, *e.g.* developers' expectations of the behavior of such tools and APIs [20]. Furthermore, it might be helpful to improve documentation for existing utilities, and to provide secure code examples [50, 57, 71]. As was shown by Jain and Lindqvist [41], providing working examples is essential for security and privacy because developers tend to follow code examples very closely.

Also, the integration of the security and privacy measures into third-party software should always be examined carefully. The survey's results, similar to the results by Kumar *et al.* [50] underline that many ML practitioners rely entirely on third-party services, not only for functionality but also for security and privacy. A certification authority that assesses these aspects would be helpful for practitioners in choosing which third-party software can be trusted. Similar assessment is already in place for other software products, such as operation systems, cryptographic modules or database servers [17].

With the support of the right tools and adequate solutions through third-party software, automated detection and defence against security and privacy threats could be advanced to further support ML practitioners. Individual awareness of this issue would then be more relevant for choosing the right tools instead of having to consider each threat and defense mechanism individually.

**Over all company sizes, there are cases, where nobody is responsible for the overall model security and privacy.** This finding is congruent with the results of a case study conducted by Flechais *et al.* [29]. During their research on supporting developers making applications more secure, they found that in many cases no one was explicitly responsible for the security of the project. The results of the present study also indicate that the practitioners responsible for ensuring ML models in their workplaces are more likely to have a higher level of education. This finding might reflect some companies' internal recruiting strategies or suggest that such high-responsibility tasks are still given to the employees with formal education proofs.

**The introduction of the GDPR had comparatively little impact on ML workflows in particular. However, there were very noticeable changes in general workflows related to personal data.** Several participants mentioned uncertainties about the formulation of the GDPR and issues with the technical feasibility for ML at some point. Such shortcomings could be addressed by having more practitioners participate in the policy-making process of data protection regulations [4]. Alternatively, the creation and provision of precise guidelines, similar to the ones that already exist for secure app development [62, 64] could prove helpful [89]. Currently, however, this represents a severe challenge, as research on privacy and security in ML is still at a stage where concrete recommendations are only possible for a relatively small subset of issues.

## 6 CONCLUSION AND FUTURE WORK

In this paper, the awareness of security and privacy among ML practitioners under several aspects was examined. Additionally, the knowledge and use of existing methods and libraries were analyzed and the impact of the GDPR on practitioners' workflows was assessed. A generally low awareness and knowledge on ML-specific privacy measures among the participants was identified, as well as a lack of institutional education on this subject. Future work could extend further in the direction of studying ML practitioners' practical workflows. This could be helpful for the design and development of more user-friendly tools and libraries to support secure and private ML. Additionally, it could serve as a basis for guidelines and regulations that take the practitioners' actual questions and needs into account.

## ACKNOWLEDGMENTS

This research work was partly funded by the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

## REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, and Zhifeng Chen *et al.* 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Yasemin Acar, Christian Stransky, Dominik Wermke, Michelle Mazurek, and Sascha Fahl. 2017. Security Developer Studies with GitHub Users: Exploring a Convenience Sample. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)* (San Antonio, Texas). USENIX Association, Santa Clara, CA, USA, 81–95.

[3] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. arXiv:1606.06565v2 http://arxiv.org/abs/1606.06565

[4] Rebecca Balebako and Lorrie Faith Cranor. 2014. Improving App Privacy: Nudging App Developers to Protect User Privacy. *IEEE Security & Privacy* 12, 4 (2014), 55–58. https://doi.org/10.1109/MSP.2014.70

[5] Rebecca Balebako, Abigail Marsh, Jialiu Lin, Jason Hong, and Lorrie Faith Cranor. 2014. The Privacy and Security Behaviors of Smartphone App Developers. In *Proceedings 2014 Workshop on Usable Security* (San Diego, CA). Internet Society, New York, NY, USA, 1–10. https://doi.org/10.14722/usec.2014.23006

[6] Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. 2010. The security of machine learning. *Machine Learning* 81, 2 (2010), 121–148. https://doi.org/10.1007/s10994-010-5188-5

[7] Marco Barreno, Blaine Nelson, Russell Sears, Anthony D. Joseph, and J. D. Tygar. 2006. Can machine learning be secure?. In *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security, ASIACCS.* ACM, New York, NY, USA, 16–25. https://doi.org/10.1145/1128817.1128824

[8] Maurice S Bartlett. 1951. The effect of standardization on a $\chi$ 2 approximation in factor analysis. *Biometrika* 38, 3/4 (1951), 337–344.

[9] Nathan Benaich and Ian Hogarth. 2019. State of AI Report 2019. https://www.stateof.ai/2019 last accessed on: Feb. 2nd 2021.

[10] Nathan Benaich and Ian Hogarth. 2020. State of AI Report 2020. https://www.stateof.ai/ last accessed on: April 7th 2021.

[11] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* 57, 1 (1995), 289–300.

[12] Patrik Berander. 2004. Using Students as Subjects in Requirements Prioritization. In *Proceedings. 2004 International Symposium on Empirical Software Engineering, 2004. ISESE '04.* IEEE, New York, NY, USA, 167–176. https://doi.org/10.1109/ISESE.2004.1334904

[13] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion Attacks against Machine Learning at Test Time. In *Advanced Information Systems Engineering*. Vol. 7908. Springer, Berlin, Heidelberg, 387–402. https://doi.org/10.1007/978-3-642-40994-3_25

[14] Battista Biggio, Blaine Nelson, and Pavel Laskov. 2012. Poisoning Attacks against Support Vector Machines. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012.* icml.cc / Omnipress, Madison, WI, USA, 1–8. http://icml.cc/2012/papers/880.pdf

[15] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. 2015. Machine Learning Classification over Encrypted Data. In *22nd Annual Network and Distributed System Security Symposium, NDSS.* The Internet Society, Reston, Virginia, USA, 14. https://www.ndss-symposium.org/ndss2015/machine-learning-classification-over-encrypted-data

[16] Brendan McMahan and Daniel Ramage. 2017. Federated Learning: Collaborative Machine Learning without Centralized Training Data. http://ai.googleblog.com/2017/04/federated-learning-collaborative.html http://ai.googleblog.com/2017/04/federated-learning-collaborative.html, last accessed on: Feb. 2nd 2021.

[17] BSI. 2020. German IT Security Certificates. https://www.bsi.bund.de/EN/Topics/Certification/certification_node.html, last accessed on: April 7th 2021.

[18] Anna L Buczak and Erhan Guven. 2015. A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications surveys & tutorials* 18, 2 (2015), 1153–1176.

[19] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and where: A characterization of data provenance. In *International conference on database theory.* [], Springer, Berlin, Heidelberg, 316–330.

[20] Justin Cappos, Yanyan Zhuang, Daniela Oliveira, Marissa Rosenthal, and Kuo-Chuan Yeh. 2014. Vulnerabilities as Blind Spots in Developer's Heuristic-Based Decision-Making Processes. In *Proceedings of the 2014 Workshop on New Security Paradigms Workshop - NSPW '14* (Victoria, British Columbia, Canada). ACM Press, New York, NY, USA, 53–62. https://doi.org/10.1145/2683467.2683472

[21] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[22] Lee J Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[23] Morten Dahl, Jason Mancuso, Yann Dupis, Ben Decoste, Morgan Giraud, Ian Livingstone, Justin Patriquin, and Gavin Uhma. 2018. Private Machine Learning in TensorFlow using Secure Computation. *CoRR* abs/1810.08130 (2018), 6. arXiv:1810.08130 http://arxiv.org/abs/1810.08130

[24] Thomas G Dietterich et al. 2002. Ensemble learning. *The handbook of brain theory and neural networks* 2 (2002), 110–125.

[25] Gavin Weiguang Ding, Luyu Wang, and Xiaomeng Jin. 2019. AdverTorch v0.1: An Adversarial Robustness Toolbox based on PyTorch. https://github.com/BorealisAI/advertorch.

[26] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Proceedings, Part II (Lecture Notes in Computer Science, Vol. 4052).* Springer, Berlin, Heidelberg, 1–12. https://doi.org/10.1007/11787006_1

[27] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science* 9, 3-4 (2014), 211–407.

[28] Michael P. Fay and Michael A. Proschan. 2010. Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules. *Statistics Surveys* 4, 0 (2010), 1–39. https://doi.org/10.1214/09-ss051

[29] Ivan Flechais, Martina Angela Sasse, and Stephen Hailes. 2003. Bringing security home: a process for developing secure and usable systems. In *Proceedings of the New Security Paradigms Workshop*, Christian Hempelmann and Victor Raskin (Eds.). ACM, New York, NY, USA, 49–57. https://doi.org/10.1145/986655.986664

[30] Joseph L. Fleiss, Bruce Levin, and Myunghee Cho Paik. 2013. *Statistical Methods for Rates and Proportions.* John Wiley & Sons, Hoboken, New Jersey, USA.

[31] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. 2015. Model Inversion Attacks That Exploit Confidence Information and Basic Countermeasures. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15* (Denver, Colorado, USA). ACM Press, New York, NY, USA, 1322–1333. https://doi.org/10.1145/2810103.2813677

[32] Galen Andrew, Steve Chien, and Nicolas Papernot. 2019. *Tensorflow/Privacy.* tensorflow. https://github.com/tensorflow/privacy https://github.com/tensorflow/privacy, last accessed on: Feb. 2nd 2021.

[33] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for Datasets. arXiv:1803.09010 http://arxiv.org/abs/1803.09010

[34] Limesurvey GmbH. 2006–2021. LimeSurvey: An Open Source Survey Tool. http://www.limesurvey.org, last accessed on: Feb. 2nd 2021.

[35] Morgane Goibert and Elvis Dohmatob. 2019. Adversarial Robustness via Adversarial Label-Smoothing. arXiv:1906.11567 http://arxiv.org/abs/1906.11567

[36] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and Harnessing Adversarial Examples. arXiv:1412.6572 [cs, stat] http://arxiv.org/abs/1412.6572

[37] Thore Graepel, Kristin Lauter, and Michael Naehrig. 2012. ML confidential: Machine learning on encrypted data. In *International Conference on Information Security and Cryptology*. Springer, Springer, Berlin, Heidelberg, 1–21.

[38] Bartlomiej Hanus, John C Windsor, and Yu Wu. 2018. Definition and multidimensionality of security awareness: Close encounters of the second order. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems* 49, SI (2018), 103–133.

[39] Martin Höst, Björn Regnell, and Claes Wohlin. 2000. Using Students as Subjects-A Comparative Study of Students and Professionals in Lead-Time Impact Assessment. *Empirical Software Engineering* 5, 3 (2000), 201–214.

[40] Ling Huang, Anthony D. Joseph, Blaine Nelson, Benjamin I. P. Rubinstein, and J. D. Tygar. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, AISec 2011, Chicago, IL, USA, October 21, 2011.* ACM, New York, NY, USA, 43–58. https://doi.org/10.1145/2046684.2046692

[41] Shubham Jain and Janne Lindqvist. 2014. Should I Protect You? Understanding Developers' Behavior to Privacy-Preserving APIs. In *Proceedings 2014 Workshop on Usable Security* (San Diego, CA). Internet Society, Reston, Virginia, USA, 10. https://doi.org/10.14722/usec.2014.23045

[42] Xin Jin and Jiawei Han. 2010. *K-Means Clustering.* Springer US, Boston, MA, 563–564. https://doi.org/10.1007/978-0-387-30164-8_425

[43] Noah Johnson, Joseph P Near, and Dawn Song. 2018. Towards practical differential privacy for SQL queries. In *Proceedings of the VLDB Endowment*, Vol. 11,5. VLDB Endowment, [], 526–539. https://github.com/uber-archive/sql-differential-privacy.

[44] Kaggle. 2020. State of Machine Learning and Data Science 2020. https://storage.googleapis.com/kaggle-media/surveys/Kaggle%20State%20of%20Machine%20Learning%20and%20Data%20Science%202020.pdf, last accessed on: April 7th 2021.

[45] Henry F Kaiser. 1970. A second generation little jiffy. *Psychometrika* 35, 4 (1970), 401–415.

[46] Henry F Kaiser and John Rice. 1974. Little jiffy, mark IV. *Educational and psychological measurement* 34, 1 (1974), 111–117.

[47] Trupti M Kodinariya and Prashant R Makwana. 2013. Review on determining number of Cluster in K-Means Clustering. *International Journal* 1, 6 (2013), 90–95.

[48] Jakub Konečnỳ, H Brendan McMahan, Felix X Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. 2016. Federated learning: Strategies for improving communication efficiency. arXiv:1610.05492 https://arxiv.org/abs/1610.05492

[49] William H. Kruskal and W. Allen Wallis. 1952. Use of Ranks in One-Criterion Variance Analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621. http://www.jstor.org/stable/2280779

[50] Ram Shankar Siva Kumar, Magnus Nyström, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. 2020. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, IEEE, New York, NY, USA, 69–75.

[51] Yehuda Lindell and Benny Pinkas. 2009. Secure Multiparty Computation for Privacy-Preserving Data Mining. *J. Priv. Confidentiality* 1, 1 (2009), 40. https://doi.org/10.29012/jpc.v1i1.566

[52] Qiang Liu, Pan Li, Wentao Zhao, Wei Cai, Shui Yu, and Victor C. M. Leung. 2018. A Survey on Security Threats and Defensive Techniques of Machine Learning: A Data Driven View. *IEEE Access* 6 (2018), 12103–12117. https://doi.org/10.1109/ACCESS.2018.2805680

[53] Ben Lorica Loukides, Mike. 2019. You Created a Machine Learning Application. Now Make Sure It's Secure. https://www.oreilly.com/ideas/you-created-a-machine-learning-application-now-make-sure-its-secure, last accessed on: Feb. 2nd 2021.

[54] H. Brendan McMahan, Galen Andrew, Ulfar Erlingsson, Steve Chien, Ilya Mironov, Nicolas Papernot, and Peter Kairouz. 2019. A General Approach to Adding Differential Privacy to Iterative Training Procedures. arXiv:arXiv:1812.06210 [cs.LG] https://github.com/tensorflow/privacy.

[55] Microsoft Research, Redmond, WA. 2019. *Microsoft SEAL (Release 3.4).* Microsoft. https://github.com/Microsoft/SEAL https://github.com/Microsoft/SEAL, last accessed on: Feb. 2nd 2021.

[56] Christoph Molnar. 2019. Interpretable Machine Learning. https://christophm.github.io/interpretable-ml-book/ https://christophm.github.io/interpretable-ml-book/, last accessed on: Feb. 2nd 2021.

[57] Sarah Nadi, Stefan Krüger, Mira Mezini, and Eric Bodden. 2016. Jumping through Hoops: Why Do Java Developers Struggle with Cryptography APIs?. In *Proceedings of the 38th International Conference on Software Engineering - ICSE '16* (Austin, Texas). ACM Press, New York, NY, USA, 935–946. https://doi.org/10.1145/2884781.2884790

[58] Yuki Nagai, Yusuke Uchida, Shigeyuki Sakazawa, and Shin'ichi Satoh. 2018. Digital watermarking for deep neural networks. *International Journal of Multimedia Information Retrieval* 7, 1 (2018), 3–16.

[59] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, Marco Herzog, Sergej Dechand, and Matthew Smith. 2017. Why Do Developers Get Password Storage Wrong?: A Qualitative Usability Study. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security - CCS '17* (Dallas, Texas, USA). ACM Press, New York, NY, USA, 311–328. https://doi.org/10.1145/3133956.3134082

[60] Alena Naiakshina, Anastasia Danilova, Christian Tiefenau, and Matthew Smith. 2018. Deception Task Design in Developer Password Studies: Exploring a Student Sample. In *Fourteenth Symposium on Usable Privacy and Security, SOUPS 2018, Baltimore, MD, USA, August 12-14, 2018*, Mary Ellen Zurko and Heather Richter Lipford (Eds.). USENIX Association, Berkeley, California, USA, 297–313. https://www.usenix.org/conference/soups2018/presentation/naiakshina

[61] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, and Ambrish Rawat *et al.* 2019. Adversarial Robustness Toolbox v1.0.0. arXiv:1807.01069 [cs.LG] https://github.com/Trusted-AI/adversarial-robustness-toolbox.

[62] Office of the Privacy Commissioner of Canada. 2012. Seizing Opportunity: Good Privacy Practices for Developing Mobile Apps. https://www.priv.gc.ca/en/privacy-topics/technology/mobile-and-digital-devices/mobile-apps/gd_app_201210/ https://www.priv.gc.ca/en/privacy-topics/technology/mobile-and-digital-devices/mobile-apps/gd_app_201210/, last accessed on: Feb. 2nd 2021.

[63] Office of the Privacy Commissioner of Canada. 2019. The Personal Information Protection and Electronic Documents Act (PIPEDA). https://www.priv.gc.ca/en/privacy-topics/privacy-laws-in-canada/the-personal-information-protection-and-electronic-documents-act-pipeda/, last accessed on: Feb. 2nd 2021.

[64] Office of the Australian Information Commissioner (OAIC). 2014. Mobile Privacy: A Better Practice Guide for Mobile App Developers. https://www.oaic.gov.au/privacy/guidance-and-advice/mobile-privacy-a-better-practice-guide-for-mobile-app-developers/, last accessed on: Feb. 2nd 2021.

[65] Stanley RM Oliveira and Osmar R Zaiane. 2003. Protecting sensitive knowledge by data sanitization. In *Third IEEE International conference on data mining*. IEEE, IEEE, New York, NY, USA, 613–616.

[66] Nicolas Papernot. 2018. A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private. In *Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security, CCS*. ACM, New York, NY, USA, 1. https://doi.org/10.1145/3270101.3270102

[67] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, and Reuben Feinman *et al.* 2018. Technical Report on the CleverHans v2.1.0 Adversarial Examples Library. arXiv:arXiv:1610.007682 https://github.com/cleverhans-lab/cleverhans.

[68] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael P Wellman. 2018. Sok: Security and privacy in machine learning. In *2018 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, IEEE, New York, NY, USA, 399–414.

[69] Nicolas Papernot and Patrick D. McDaniel. 2018. Deep K-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. arXiv:1803.04765 http://arxiv.org/abs/1803.04765

[70] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, and Thirion *et al.* 2011. Scikit-learn: Machine learning in Python. *the Journal of machine Learning research* 12 (2011), 2825–2830.

[71] Olgierd Pieczul, Simon Foley, and Mary Ellen Zurko. 2017. Developer-Centered Security and the Symmetry of Ignorance. In *Proceedings of the 2017 New Security Paradigms Workshop on ZZZ - NSPW 2017* (Santa Cruz, CA, USA). ACM Press, New York, NY, USA, 46–56. https://doi.org/10.1145/3171533.3171539

[72] Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. 2020. Foolbox Native: Fast adversarial attacks to benchmark the robustness of machine learning models in PyTorch, TensorFlow, and JAX. In *Journal of Open Source Software*, Vol. 5, 53. The Open Journal, [], 2607. https://doi.org/10.21105/joss.02607 https://github.com/bethgelab/foolbox.

[73] Theo Ryffel, Andrew Trask, Morten Dahl, Bobby Wagner, Jason Mancuso, Daniel Rueckert, and Jonathan Passerat-Palmbach. 2018. A generic framework for privacy preserving deep learning. arXiv:1811.04017 [cs.LG] https://github.com/OpenMined/PySyft.

[74] Iflaah Salman, Ayse Tosun Misirli, and Natalia Juristo. 2015. Are Students Representatives of Professionals in Software Engineering Experiments?. In *2015 IEEE/ACM 37th IEEE International Conference on Software Engineering*, Vol. 1. IEEE, New York, NY, USA, 666–676. https://doi.org/10.1109/ICSE.2015.82

[75] Ravi S Sandhu and Pierangela Samarati. 1994. Access control: principle and practice. *IEEE communications magazine* 32, 9 (1994), 40–48.

[76] Educational Testing Service. 2019. factor_analyzer: Open source Python module to perform exploratory and factor analysis. https://factor-analyzer.readthedocs.io/en/latest/index.html/.

[77] Yi Shi, Yalin E Sagduyu, Kemal Davaslioglu, and Jason H Li. 2018. Active deep learning attacks under strict rate limitations for online API calls. In *2018 IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE, IEEE, New York, NY, USA, 1–6.

[78] Justin Smith, Brittany Johnson, Emerson Murphy-Hill, Bill Chu, and Heather Richter Lipford. 2015. Questions Developers Ask While Diagnosing Potential Security Vulnerabilities with Static Analysis. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering - ESEC/FSE 2015* (Bergamo, Italy). ACM Press, New York, NY, USA, 248–259. https://doi.org/10.1145/2786805.2786812

[79] Mark Stamp. 2011. *Information security: principles and practice*. John Wiley & Sons, Hoboken, New Jersey, USA.

[80] State of California Department of Justice. 2018. California Consumer Privacy Act (CCPA). https://oag.ca.gov/privacy/ccpa https://oag.ca.gov/privacy/ccpa, last accessed on: Feb. 2nd 2021.

[81] Jeffrey Stylos and Brad A. Myers. 2008. The implications of method placement on API learnability. In *Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering, 2008, Atlanta, Georgia, USA, November 9-14, 2008*, Mary Jean Harrold and Gail C. Murphy (Eds.). ACM, New York, NY, USA, 105–112. https://doi.org/10.1145/1453101.1453117

[82] Vinith M Suriyakumar, Nicolas Papernot, Anna Goldenberg, and Marzyeh Ghassemi. 2021. Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA, 723–734.

[83] Mikael Svahnberg, Aybüke Aurum, and Claes Wohlin. 2008. Using students as subjects - an empirical evaluation. In *Proceedings of the Second International Symposium on Empirical Software Engineering and Measurement, ESEM*, H. Dieter Rombach, Sebastian G. Elbaum, and Jürgen Münch (Eds.). ACM, New York, NY, USA, 288–290. https://doi.org/10.1145/1414004.1414055

[84] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. 2017. Embedding Watermarks into Deep Neural Networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR*, Bogdan Ionescu, Nicu Sebe, Jiashi Feng, Martha A. Larson, Rainer Lienhart, and Cees Snoek (Eds.). ACM, New York, NY, USA, 269–277. https://doi.org/10.1145/3078971.3078974

[85] Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

[86] Dinusha Vatsalan, Peter Christen, and Vassilios S. Verykios. 2013. A taxonomy of privacy-preserving record linkage techniques. *Inf. Syst.* 38, 6 (2013), 946–969. https://doi.org/10.1016/j.is.2012.11.005

[87] Dinusha Vatsalan, Peter Christen, and Vassilios S Verykios. 2013. A taxonomy of privacy-preserving record linkage techniques. *Information Systems* 38, 6 (2013), 946–969.

[88] Michael Veale, Reuben Binns, and Lilian Edwards. 2018. Algorithms that Remember: Model Inversion Attacks and Data Protection Law. *CoRR* abs/1807.04644 (2018), 15. arXiv:1807.04644 http://arxiv.org/abs/1807.04644

[89] Denis Verdon. 2006. Security policies and the software developer. *IEEE Security & Privacy* 4, 4 (2006), 42–49. https://doi.org/10.1109/MSP.2006.103

[90] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, and SciPy 1.0 Contributors *et al.* 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods* 17 (2020), 261–272. https://doi.org/10.1038/s41592-019-0686-2

[91] Paul Voigt and Axel von dem Bussche. 2017. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer, Berlin, Heidelberg.

[92] Christina Voskoglou. 2017. What is the best programming language for Machine Learning? https://towardsdatascience.com/what-is-the-best-programming-language-for-machine-learning-a745c156d6b7, last accessed on: April 7th 2021.

[93] Christopher Waites. 2019. PyVacy: Towards Practical Differential Privacy for Deep Learning. http://hdl.handle.net/1853/61412, last accessed on: Feb. 2nd 2021.

[94] Royce J Wilson, Celia Yuxin Zhang, William Lam, Damien Desfontaines, Daniel Simmons-Marengo, and Bryant Gipson. 2019. Differentially Private SQL with Bounded User Contribution. arXiv:1909.01917 [cs.CR] https://github.com/google/differential-privacy.

[95] Glenn Wurster and P. C. van Oorschot. 2008. The Developer Is the Enemy. In *Proceedings of the 2008 Workshop on New Security Paradigms - NSPW '08* (Lake Tahoe, California, USA). ACM Press, New York, NY, USA, 89. https://doi.org/10.1145/1595676.1595691

[96] P.J Zarco-Tejada, C.A Rueda, and S.L Ustin. 2003. Water Content Estimation in Vegetation with MODIS Reflectance Data and Model Inversion Methods. *Remote Sensing of Environment* 85, 1 (2003), 109–124. https://doi.org/10.1016/S0034-4257(02)00197-9

## APPENDIX

Table 1. List of selected items for the factor analysis to be applied on. Only items with loadings > 0.45 were considered to correlate strong enough with the latent construct resulting in five variables being excluded from further analysis — namely *Adversarial Training*, *Observing model input at inference time*, *Smoothing prediction output*, *Federated Learning and System Security* — for which respective loadings are not reported.

| Item | Factor 1 |
|---|---|
| Inversion Attack | .669 |
| Impersonation Attack | .588 |
| Poisoning Attack | .663 |
| Evasion Attack | .654 |
| Data Sanitization | .623 |
| Data Provenance | .543 |
| Adversarial Training | |
| Ensemble Learning | .481 |
| Observing model input at inference time | |
| Smoothing prediction output | |
| Federated Learning | |
| Introducing delay for model interaction | .543 |
| Access Control | .532 |
| System Security | |
| Differential Privacy | .548 |
| Homomorphic Encryption | .539 |
| Watermarking | .503 |
| Privacy-preserving Record Linkage | .644 |
| % of total variance | 33.94 |
| Cronbach's Alpha | 0.86 |

Table 2. Summary of participants' background information. Demographics marked with an * allowed for multiple answers.

| | |
|---|---|
| **Area** | |
| Europe | 64 |
| North America | 9 |
| South America | 2 |
| Asia | 5 |
| Australia | 3 |
| **Education** | |
| High school / Secondary school degree | 2 |
| Bachelor's degree | 12 |
| Master's degree | 55 |
| Doctorate | 13 |
| Other | 1 |
| **Employment** | |
| Employed | 73 |
| Self-employed | 6 |
| Unemployed | 4 |
| **ML Application*** | |
| Industry | 38 |
| Industrial Research | 39 |
| Academic Research | 48 |
| Hobby | 17 |
| **Working experience in ML** | |
| 1-3 years | 49 |
| 4-6 years | 18 |
| 7-9 years | 5 |
| 10 years or more | 11 |
| **Company size (# of employees)** | |
| Self-employed | 6 |
| 1-10 | 5 |
| 11-50 | 8 |
| 51-200 | 10 |
| 201-500 | 12 |
| 501-1000 | 8 |
| 1001-5000 | 15 |
| 5001-10 000 | 7 |
| More than 10 000 | 12 |

Table 3. Summary of participants' ML working environment. Demographics marked with an * allowed for multiple answers and, therefore, do not add up to the sample size of 83.

| | |
|---|---|
| **Domain(s) of the ML data*** | |
| Customers and users | 32 |
| Smart environment and IoT | 22 |
| Medical and health | 20 |
| Transportation and traffic | 19 |
| Financial | 19 |
| Public security | 11 |
| Weather and environment | 8 |
| Social media | 7 |
| Other | 22 |
| **Type(s) of data handled*** | |
| Images | 44 |
| Sensor data | 39 |
| Tabular data | 38 |
| Text | 37 |
| Metadata | 30 |
| Location data | 24 |
| Video | 17 |
| Audio/Sound | 15 |
| Other | 4 |
| **Daily ML Task(s)*** | |
| Applying ML libraries | 59 |
| Data analysis | 54 |
| Evaluation | 47 |
| Data cleansing and preparation | 45 |
| Coordinating ML projects and workflows | 44 |
| Developing custom ML applications | 36 |
| Data collection | 31 |
| Deployment and maintenance | 27 |
| Developing ML tools / libraries from scratch | 16 |
| **Role of ML in product development** | |
| Key part of the product | 47 |
| Included in the product but not key part | 28 |
| Used internally for other than marketing | 6 |
| Used internally only for marketing | 2 |

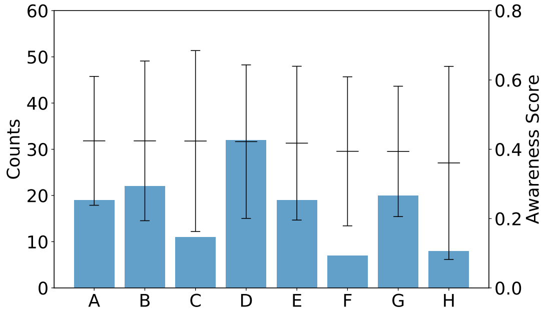


Fig. 11. Absolute number of participants per data domain in connection with the average awareness value and standard deviation resulting in the respective field. A: 'Financial', B: 'Smart Env. and IOT', C: 'Public Security', D: 'Customers and Users', E: 'Transport and Traffic', F: 'Social Media', G: 'Medical and Health', H: 'Weather and Environment'.

Table 4. Set of hypotheses to answer the research question RQ1. All hypotheses in this table form a hypothesis family and are considered as such in the process of correction for multiple comparisons.

| No. | Hypotheses | Test statistic | p-value | $p^*$ |
|---|---|---|---|---|
| H.1.1 | $H_0$: Awareness does not correlate with the individual's perception of how important ML security is. | $r_{(81)} = .073$ | .514 | .62 |
| H.1.2 | $H_0$: Educational degree has no effect on the participants' level of awareness. | $\chi^2_{(4)} = .89$ | .827 | .827 |
| H.1.3 | $H_0$: Awareness does not correlate with years of working experience in ML. | $r_{(81)} = .36$ | $< .001$ | .005 |
| H.1.4 | $H_0$: Company size has no effect on the participants' level of awareness. | $\chi^2_{(8)} = 15.26$ | .054 | .109 |
| H.1.5 | $H_0$: Participants who build their own ML applications do not have a higher awareness than those who don't. | $U_{(41,42)} = 584.0$ | .006 | .018 |
| H.1.6 | $H_0$: There is no difference in the level of awareness between participants working with directly human-related data, indirectly or data that is not human-related at all. | $\chi^2_{(2)} = 4.09$ | .129 | .194 |

Table 5. Set of hypotheses to answer the research question RQ2. All hypotheses in this table form a hypothesis family and are considered as such in the process of correction for multiple comparisons.

| No. | Hypotheses | Test statistic | p-value | $p^*$ |
|---|---|---|---|---|
| H.2.1 | $H_0$: Years of working experience do not correlate with the individual's perception of how important ML security is. | $r_{(81)} = .635$ | .57 | .685 |
| H.2.2 | $H_0$: Company size has no effect on who is responsible for securing ML models in the respective working environment. | $\chi^2_{(32)} = 43.11$ | .091 | .182 |
| H.2.3 | $H_0$: Educational degree has no effect on whether someone is responsible for implementing security solutions. | $U_{(47,35)} = 774.5$ | .296 | .395 |
| H.2.4 | $H_0$: The number of years of working experience with ML has no effect on whether someone is responsible for implementing security solutions. | $U_{(48,35)} = 728.5$ | .124 | .212 |
| H.2.5 | $H_0$: The introduction of the GDPR had no effect on the individuals' ML security practices - grouped by how human-related the data used is. | $\chi^2_{(2)} = 12.39$ | .002 | .005 |
| H.2.5.1 | $H_0$: The introduction of the GDPR had no effect on the individuals' ML security practices with respect to those working with non-human related data and those who work with indirectly related data. | $U_{(23,17)} = 97.0$ | .002 | .005 |
| H.2.5.2 | $H_0$: The introduction of the GDPR had no effect on the individuals' ML security practices with respect to those working with non-human related data and those who work with directly related data. | $U_{(23,23)} = 130.0$ | $< .001$ | .003 |
| H.2.5.3 | $H_0$: The introduction of the GDPR had no effect on the individuals' ML security practices with respect to those working with indirectly related data and those who work with directly related data. | $U_{(17,23)} = 171.5$ | .254 | .38 |
| H.2.6 | $H_0$: The four attacks described are all equally well known. | $\chi^2_{(3)} = 1.12$ | .772 | .772 |
| H.2.7 | $H_0$: The four attacks described are all equally often implemented. | $\chi^2_{(3)} = 1.31$ | .727 | .772 |
| H.2.8 | $H_0$: The 14 methods described are all equally well known. | $\chi^2_{(13)} = 93.19$ | $< .001$ | $< .001$ |
| H.2.9 | $H_0$: The 14 methods described are all equally often implemented. | $\chi^2_{(13)} = 209.0$ | $< .001$ | $< .001$ |

# Security and Privacy in Machine Learning

This survey aims to analyze the state of the art in implementation of security measures to protect machine learning (ML) systems. We want to investigate to what extent machine learning practitioners are aware of the different types of attacks (from inside and outside) that their models are exposed to. We also want to learn which types of protective measures are used. Finally, we are interested in learning what kind of experiences developers have had with fulfilling the requirements of the European General Data Protection Regulation (GDPR).

We kindly ask all respondents to stick to the truth as best as they can and avoid exaggerating their statements in any direction, as this ensures the validity of the results. This survey will take approximately 10-15 minutes to fill out and consists of a maximum of 25 questions.

The survey is conducted by  is conducted by the Fraunhofer Institute for Applied and Integrated Security (AISEC)  in cooperation with the Fraunhofer Institute for Secure Information Technology (SIT) and the Freie Universität Berlin (FU).

- ❑ Indicate questions that allow for multiple answers.
- ○ Indicate questions that allow for exactly one answer.

*Data security and consent note:*

*Participation in this study is voluntary. You can discontinue your participation at any time with no negative consequences, but information gathered from you up until the point of cessation of your participation may be used in the study. The data collected within this study include questionnaire items regarding your experience, awareness, and knowledge. These data can not be linked to your person. The research data are collected purely for scientific purposes. The research data are only available to the researchers of the research group. The research team deploys appropriate technical and organizational security precautions to protect personal data against disappearance, misuse, unlawful use, change, or destruction. The data collected of you within this survey are retained as long as is necessary for the purpose it should fulfil, or as long as the legislation requires. Any contact information we might collect of you is separated from other questionnaire data, including demographic information. Upon request you are provided with additional details of the general principles of this study and its progress, or of the results concerning yourself.*

- ○ Agree
- ○ Disagree

[Only participants who gave their consent were forwarded to the questionnaire.]

# Demographics

*1 Which country are you currently working in?*

○ Afghanistan
...
○ Zimbabwe

*2 What is the highest educational degree you have obtained?*

○ Less than high school or secondary school degree (i.e. Abitur, baccalauréat, A levels etc.)
○ High school or secondary school degree
○ Bachelor's degree
○ Master's degree or diploma
○ Doctorate
○ Other: [ ]

*3 What is your current working situation?*

○ I am a student. (If you are also working in a machine learning related position at the same time, please check employee or self-employed, according to what applies to you!)
○ I am an employee.
○ I am self-employed.
○ I am unemployed.

*4 How long have you been working with machine learning (ML) - either professionally or as a hobby?*

○ I have never worked with ML
○ 1-3 years
○ 4-6 years
○ 7-9 years
○ 10 years or more

*5 What are your daily machine learning (ML)-related tasks?*

□ Coordinating ML projects and workflows
□ Applying ML libraries (tensorflow, scikit learn, ...)
□ Developing custom ML applications (e.g. design custom neural networks for given tasks)
□ Developing ML tools or libraries from scratch
□ Data cleansing and preparation
□ Data analysis
□ Data collection
□ Evaluation
□ Deployment and maintenance
□ Other: [ ]

*6 For what field(s) do you apply machine learning?*

❑ Industry
❑ Industrial research
❑ Academic research
❑ Hobby
❑ Other: [ ]

*7 What is the size of the company you are working for?*

** This question was not displayed for participants who had indicated being a student.*

○ Self-employed
○ 1-10 employees
○ 11-50 employees
○ 51-200 employees
○ 201-500 employees
○ 501-1000 employees
○ 1001-5000 employees
○ 5001-10,000 employees
○ 10,001+ employees

*8 How are the product(s) that your division develops concerned with machine learning (ML)?*

** This question was not displayed for participants who had indicated being a student.*

○ ML is key part of the product.
○ ML is included in the product but not key part.
○ ML is only used internally for marketing.
○ ML is only used internally for other purposes than marketing (e.g. to improve the product, finance, ...).

## Data and Sensitivity

*9 Is any of the data you work with sensitive?*

Sensitive data means information that has to be be protected against unwarranted disclosure (e.g. private or confidential data).

○ No
○ Yes

*10 Do your machine learning models deal with data of individiuals?*

I work with data that is...

○ ...not related to humans (any of my data).
○ ...indirectly related to humans (at least some data that I work with).
○ ...directly related to humans (at least some data that I work with).

*11 What type of data do you deal with in your machine learning models?*

❑ Images
❑ Video
❑ Audio/Sound
❑ Text
❑ Location data
❑ Metadata
❑ Sensor data
❑ Tabular data
❑ Other: [ ]

*12 What domain does the data you are working with stem from?*

❑ Financial
❑ Medical and health
❑ Transportation and traffic
❑ Customers and users
❑ Weather and environment
❑ Smart environment and IoT
❑ Social media
❑ Public security
❑ Other: [ ]

# ML Security

*13 In your opinion, how important or unimportant is it to ensure the security of your machine learning models?*

○ Unimportant
○ Of little importance
○ Moderately important
○ Important
○ Very important

*14 How did you build your current knowledge about machine learning security?*

❑ Through courses at university
❑ Through workshops and tutorials
❑ Through practice
❑ Through self-study (e.g. online tutorials, webinars, literature)

- □ Other: [ ]

*15 Who takes care of the security of the machine learning (ML) models in your working environment? (If you are unemployed, please check the answer that applies to your previous job.)*

** This question was not displayed for participants who had indicated being a student.*

- ○ I take care of my ML projects' security.
- ○ I solely take care of all ML security.
- ○ I take care of all ML security, together with some others.
- ○ A designated expert takes care of ML security.
- ○ Nobody takes care of ML security.
- ○ Make a comment on your choice here: [ ]

# Attacks on ML

*16 For the following attacks against machine learning (ML), please check what applies to you.*

| | Yes, I have implemented solutions to prevent this type of attack. | I am familiar with this type of attack, but have not implemented solutions against it, yet. | No, I am not familiar with this type of attack. |
|---|---|---|---|
| Inversion attacks<br>Inversion attack: The aim of an inversion attack is to extract information from your ML model. An attacker could query your model to obtain knowledge about the underlying training data. | ○ | ○ | ○ |
| Impersonation attacks<br>Impersonation attack: To impersonate an individual from your dataset, attackers try to imitate data records of their victims. They can use those records to get unauthorized access, or to develop specially tailored attacks against that victim. | ○ | ○ | ○ |
| Poisoning attacks<br>Poisoning attack: During training, an attacker is able to inject their own data records into your training data. Your model might thereby learn things it is not supposed to, due to the shift of | ○ | ○ | ○ |

| | | | |
|---|---|---|---|
| classification boundaries. This could be exploited by the attacker in the prediction phase. | | | |
| Evasion attacks<br>Evasion attack: At test time, an attacker modifies a data record in such a minimal way, that the record still seems normal to a human observer. The modification however causes your ML model to make a prediction that differs completely from the one on the original input. Adversarial examples are an instance of evasion attacks. | o | o | o |

## ML Security Practices

*17 For the following libraries, related to private and secure machine learning (ML), please select what applies to you.*

| | I have already worked with this library. | I have heard about this library but have not used it, yet. | I have never heard about this library before. |
|---|---|---|---|
| Tensorflow Privacy | o | o | o |
| Cleverhans | o | o | o |
| PySyft | o | o | o |
| Google's Differential Privacy | o | o | o |
| Uber SQL Differential Privacy | o | o | o |
| AdverTorch | o | o | o |
| Foolbox | o | o | o |
| Adversarial Robustness Toolbox (ART) | o | o | o |

*18 Have you ever implemented a method for...*

| | Yes, I have implemented this method. | I am familiar with this method, but have not implemented it, yet. | No, I am not familiar with this method. |
|---|---|---|---|
| ...data sanitization? | o | o | o |

| | | | |
|---|---|---|---|
| Data sanitizition: All your training data is cleaned from potentially malicious data points. Samples that have a negative impact on the model's prediction output might be discarded. | | | |
| ...data provenance?<br>Data provenance: For all your training data, the provenance is clear and traceable. Your data pipeline and data storages are well documented and protected against intrusions. | o | o | o |
| ...adversarial training?<br>Adversarial training: Your model is trained partly on adversarial samples with corresponding labels to detect them as such and react adequately. | o | o | o |
| ...ensemble learning to make your ML models more secure?<br>Ensemble learning: You group several ML models into an assembly for your predictions. Hereby, different classifiers or different techniques for defence can be combined to mitigate the success of attacks and to make the model more robust. | o | o | o |
| ...observing model input at inference time?<br>Observing model input at inference time: You are observing the data that is presented to your model when it is deployed. ML models are most likely to fail when the data distribution at test time differs from the one at training time. By observing the input to your model, you can prevent an attacker using this fact to his advantage. | o | o | o |
| ...smoothing prediction output?<br>Smoothing prediction output: By rounding or truncating the prediction output slightly, or preventing sensitive outputs, you make it more difficult for an attacker to reconstruct the model or to invert it. | o | o | o |
| ...federated learning (FL)? | o | o | o |

| | | | |
|---|---|---|---|
| Federated Learning: FL is an ML technique in which the model is trained across multiple decentralized devices or parties on their local data samples. In contrast to traditional ML techniques, where all data samples are uploaded to one central server for training. In FL, no data samples are exchanged. | | | |

*19 Have you ever implemented a method for...*

| | Yes, I have implemented this method. | I am familiar with this method, but have not implemented it, yet. | No, I am not familiar with this method. |
|---|---|---|---|
| …introducing delay for model interaction?<br>Introducing delay for model interacton: You do not allow unlimitedly many and unlimitedly frequent queries to your model by e.g. introducing a delay in your responses. Thereby, it gets more difficult for the attacker to build his own copy of your model that he can exploit or alter in order to harm you. | o | o | o |
| …access control to protect your ML models?<br>Access control: You ensure that each instance that interacts with your model has only the access necessary to perform its tasks. This can also include not giving the learned model access to the training data, once that training is completed. | o | o | o |
| …system security to protect your ML models?<br>System security: You deploy your ML models on secure servers and protect certain hardware components, such as GPUs, TPUs etc., against attacks. | o | o | o |
| ...differential privacy (DP)?<br>Differential privacy: DP gives strong mathematical guarantees on the privacy of the data that you are using. It assumes an attacker with maximal knowledge and provides an upper bound on possible privacy breaches. | o | o | o |

| | | | |
|---|---|---|---|
| ...homomorphic encryption (HE)?<br>Homomorphic encryption: HE allows to perform arithmetic operations directly on the encrypted data without having to convert it into plain text first. Each operation provides an encrypted result, which, when decrypted, corresponds to the result that would have been obtained if the operation had been performed on the unencrypted data. | o | o | o |
| ...watermarking?<br>Watermarking: You poison the training data of your model yourself in order to have your model react to certain (secret) triggers. This can, amongst others, help to identify stolen copies of your model and protect your intellectual property. | o | o | o |
| ...privacy preserving record linkage?<br>Privacy Preserving Record Linkage: Some attributes of your data, that individually do not seem too sensitive, can act as pseudo-identifier for a data point, when considered together. Privacy preserving record linkage transforms this weaknes into a strength, by calculating hash values over those values, so that data across multiple datasets can be shared by different parties without disclosing the sensitive attributes. | o | o | o |

# GDPR

*20 How familiar are you with the requirements that the EU's General Data Privacy Regulation (GDPR) places on the handling of personal data?*

○ Not at all familiar
○ Slightly familiar
○ Moderately familiar
○ Familiar
○ Extremely familiar

*21 In your work with machine learning, have you been dealing with the fulfilment of GDPR requirements?*

○ Yes
○ I don't know
○ No

*22 To what extent has the GDPR caused a change in your machine learning security practices?*

○ Not at all
○ Very little
○ Somewhat
○ Very much
○ To a great extent

*23 What kind of changes has the GDPR prompted in your machine learning security practises? Please describe in a few words or sentences.*

Please write your answer here: [ ]

# Final

*24 If you would like to participate in a potential follow-up study, please enter your email address.*

Please write your answer here: [ ]

*25 If you would like to be informed about any publications resulting from this survey, please enter your email address.*

Please write your answer here: [ ]

We highly appreciate the time you took to fill out this questionnaire. Your contribution supports the research community in advancing the field of security for machine learning!

Sincerely, your Fraunhofer AISEC, SIT and FU Berlin team.

For any further questions, please do not hesitate to contact us via: securemachinelearning[at]aisec.fraunhofer.de. Your answers have been transmitted, you can close this window now.