2019 IEEE International Conference on Big Data (Big Data)

Privacy and Security of Big Data in AI Systems: A Research and Standards Perspective

Saharnaz Dilmaghani[1], Matthias R. Brust[1], Grégoire Danoy[1,2], Natalia Cassagnes[3], Johnatan Pecero[3], Pascal Bouvry[1,2]

[1]Interdisciplinary Centre for Security, Reliability, and Trust (SnT), University of Luxembourg
[2]Faculty of Science, Technology and Communication (FSTC), University of Luxembourg
[3]Agence pour la Normalisation et l'Économie de la Connaissance (ANEC G.I.E.), Luxembourg

***Abstract*—The huge volume, variety, and velocity of big data have empowered Machine Learning (ML) techniques and Artificial Intelligence (AI) systems. However, the vast portion of data used to train AI systems is sensitive information. Hence, any vulnerability has a potentially disastrous impact on privacy aspects and security issues. Nevertheless, the increased demands for high-quality AI from governments and companies require the utilization of big data in the systems. Several studies have highlighted the threats of big data on different platforms and the countermeasures to reduce the risks caused by attacks. In this paper, we provide an overview of the existing threats which violate privacy aspects and security issues inflicted by big data as a primary driving force within the AI/ML workflow. We define an adversarial model to investigate the attacks. Additionally, we analyze and summarize the defense strategies and countermeasures of these attacks. Furthermore, due to the impact of AI systems in the market and the vast majority of business sectors, we also investigate Standards Developing Organizations (SDOs) that are actively involved in providing guidelines to protect the privacy and ensure the security of big data and AI systems. Our far-reaching goal is to bridge the research and standardization frame to increase the consistency and efficiency of AI systems developments guaranteeing customer satisfaction while transferring a high degree of trustworthiness.**

I. Introduction

The huge volume of data generated by various sources, from connected devices to social media, termed as big data [1], is a valuable asset. The availability and widespread applications of big data [2] significantly impacts the growth of Machine Learning (ML) and Artificial Intelligence (AI) with the goals of increasing the efficiency and the accuracy of prediction and decision making and also minimizing their computational cost. Statistics depict the interest of the world market in AI systems that, only between 2018 and 2019, has increased by 154%, reached a $14.7 billion market size and will reach almost $37 billion by 2025 [3]. Stakeholders such as governments and industry sectors are attracted to benefit from AI to acquire insights from the data for customized services depend on customer's needs.

The integration of AI in various domains [4] significantly increases concerns regarding the privacy and security of data. The data that actuates AI includes various sensitive information, particularly individuals' information, including: images, speech, comments and posts on social media [5], [6], financial transactions, and health record information. Feeding such data in AI systems, they become vulnerable to privacy and security attacks that are even significantly increased recently [7], [8]. In a recent paper [7], the impact of adversarial attacks against AI medical systems is described such that an image of a benign melanocytic nevus is recognized as malignant with a high confidence score. A malicious attack on a face recognition system can reveal individuals images which are used to train the system [9]. By abusing a speech recognition system, an adversary can produce almost the same voice, however, transcribed the phrases [10]. Other attack techniques can cause potential safety hazards by effectively fooling the image classification system of an autonomous vehicle [11]. IoT devices as one of the major sources of big data have caused new adversarial opportunities against privacy and data protection of AI systems which are addressed in our previous work [12].

The demand for AI in the market and yet the vulnerability of the data in the workflow has stimulated Standards Developing Organizations (SDOs) to set up Subcommittees (SCs) and initiate projects [13], [14] with the mandate of providing standards and guidelines for big data and AI in order to help business sectors and market for a secure AI adoption. The Joint Technical Committee between the International Organization for Standardization and International Electrotechnical Commission (ISO/IEC JTC 1) [1] is a pioneer organization that is currently involved in developing standards on big data and AI.

Different surveys in the literature have followed a particular perspective to tackle the privacy and security of machine learning and AI systems. Bae et al. [15] have considered the vulnerabilities of AI systems in the *white-box/black-box* scenarios, while Liu et al. [16] focused on learning techniques and classified attacks based on *training/testing* phases. Biggio et al. [8] proposed a four-dimensional model based on the *goal*, *knowledge*, *capability* and *attacking strategy* of the adversary. Additionally, in [17] the authors focused on the privacy and security issues of big data from another perspective based on the three main phases of big data analysis: *data preparation*, *data processing*, and *data analysis*. In an ongoing project by ISO/IEC JTC 1, the threats against the trustworthiness of AI

[1]https://www.iso.org/isoiec-jtc-1.html

978-1-7281-0858-2/19/$31.00 ©2019 IEEE

systems are summarized and the characteristics of each has been reported [18]. We focus on the data violation threats in AI systems which are highlighted the most in the literature [8], [15], [16] and standardization [18].

This study provides a literature review on privacy and security issues of big data in AI systems. Our work departs from previous studies by discussing this issue using standards and guidelines developed by SDOs. Due to the worldwide importance of big data and AI in the market, we aim both research and standards to emphasize the opportunities where both frames can benefit from the outcomes of the other.

The remainder of this paper is organized as follows. Section II describes the concepts that are used throughout the paper. Section III presents an overview of existing studies and standards from the privacy and security of big data in AI systems. Section IV explains the countermeasures and defense mechanisms for the attacks described in Section III. Finally, Section V summarizes the paper by discussing the outcomes and insights.

II. Background

A. Machine Learning (ML) and Artificial Intelligence (AI)

In computer science, AI is associated with the accomplishments of tasks or problems by computers for which human intelligence is assumed to be required. AI is designed such that the input is the information acquired from the environment and takes actions to maximize success in achieving particular goals [19]. The most dominant way of achieving AI nowadays is by Machine Learning (ML) techniques which are build based on the concept of "without being explicitly programmed". In principle, ML consists of a set of algorithms and statistical models for computer systems to efficiently perform a particular task without relying on rule-based programming or human interaction. Developing the mathematical model is strongly dependant on the dataset, referred to as training data, which allows the program to gradually improve through the experiences and learning process from the data for predicting, detecting or making decisions [20]. A standard terminology of AI and Big data is also described in a standard document [21], an under development project from ISO/IEC JTC 1.

Machine learning techniques can be classified in different ways. In an underdevelopment standard [22] a set of ML approaches are defined as follows:

1) Supervised learning,
2) Unsupervised learning,
3) Semi-supervised learning,
4) Reinforcement learning,
5) Transfer learning.

Several techniques exist in each approach which are used based on the learning purpose and dataset. Regression, for instance is one of the well-known techniques used for prediction on labeled dataset. Clustering is another fundamental technique that is implied on unlabeled dataset for various applications such as recommendation of new options. However, clustering results shown to be highly influenced by the underlying data structure [23], [24]. Hence, a small change implied by an adversary can affect the results in the favour of the adversary [11].

B. Adversarial Model

We investigate privacy and security attacks of big data in AI systems that are modeled based on ML techniques. Each step in the workflow of the AI system can be the target of the specific attack(s). Hence, we use four phases in the AI overflow to identify the attacks based on the phase that an adversary penetrates to violate the system. The phases are illustrated in Fig. 1 on the defined AI workflow system. The first phase, *Training phase*, is the step where the trained data is fed into the ML model for the learning process. The data in this stage (labeled or unlabeled) is a significantly valuable source for the AI system that can be the aim of many attackers to violate the privacy and security [9], [25]–[27]. The next phase is the *Model phase* where the ML algorithm learns from the trained dataset and develops a model, which is the other valuable intellectual property of AI systems and hence is the target of various attacks [9], [28], [29]. The novel data is then fed into the trained model, named as *Apply phase*, where an adversary can penetrate the system and modify the results in his favor [11], [30], [31]. Finally, the valuable outcomes of the system, determined as the *Inference* phase, may host attacks that disclose sensitive information [32], [32], [33].

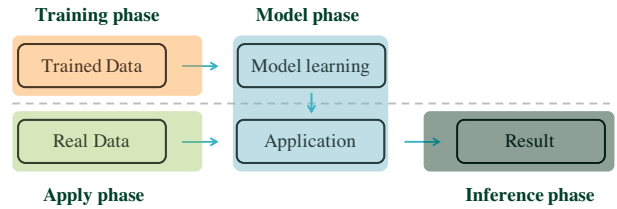

Fig. 1: The workflow and different phases of AI systems developed based on ML algorithms.

The security goals of the attacks are also investigated as the other feature. For this purpose, we consider the CIA triad [34], as the three pillars to cover the security of a system. They are summarized as follows:

- *Confidentiality* ensures the protection of sensitive information against misuse and unauthorized access. Hence, it roughly represents the privacy of a system.
- *Integrity* refers to the consistency and accuracy of data through the AI system workflow against unauthorized modification. An attack may modify the system towards misclassification, and yet does not affect the performance of the systems.
- *Availability* describes the system power to perform to achieve the expected purpose designed for the AI system with reliable outputs.

C. Standards Developing Organization (SDO)

SDOs develop technical standards and guidelines to address the needs and demands of particular adopters. Moreover,

the standards play an important role in achieving interoperability and portability of complex ICT technologies and platforms. They can bring significant benefits to industry and consumers. The best known SDO is an International Organization for Standardization (ISO) that together with International Electrotechnical Commission (IEC) initiated a Joint Technical Committee (JTC) for Information technology, known as ISO/IEC JTC 1. It covers several domains concerning smart ICT and information technology including privacy, data protection and security of ICT technologies mainly under Subcommittee, ISO/IEC JTC1/SC 27 – "Information Security, Cybersecurity and Privacy Protection", and ISO/IEC JTC1/SC 42 – "Artificial Intelligence" that is dedicated to AI that is recently created and dedicated to AI and big data. Overall, the JTC 1 has already published more than 3k standards in different domains regarding smart ICT, among them are 188 for SC 27, 3 for SC 42 with 13 more standards under development for AI and big data. The other international level SDO is the International Telecommunication Union's Telecommunication Standardization Sector (ITU) that is focused on the AI in communication technologies. Furthermore, the Institute of Electrical and Electronics Engineers (IEEE) as the other international leading standard body, has also initiated projects which mostly concern the legal and ethical perspectives of AI [13].

In the European level, European Committee for Standardization (CEN) and European Committee for Electrotechnical Standardization (CELENEC) have recently announced [14] the development of "Focus Group Artificial Intelligence" to develop standardization road-map for AI according to the European requirements. Moreover, European Telecommunications Standards Institute (ETSI) has also initiated projects focused on the use cases, applications and security challenges of AI. In this paper, our main target is the joint committee of ISO/IEC JTC 1 since it has already established a particular committee and various study and working groups in AI and big data related issues.

TABLE I: Identifying the phases where a particular attack penetrates the AI system.

Attack	**AI Workflow Phase**			
	Training	**Model**	**Apply**	**Inference**
Data Breach	✔	✔		✔
Bias in Data	✔			
Data Poisoning	✔			
Model extraction		✔		
Evasion			✔	

III. Privacy and Security of Big Data in AI

In this section, we analyze the data privacy and security attacks concerning the defined characteristics (cf. Section I). We describe the phase where the attack is imposed, the risks caused by the attack, and the real-world attack examples. Besides, an overview of the research papers and standards is conducted corresponding to each attack scenario. Table I represents, for each attack the phase(s) where a particular attack penetrates the AI system. Table II summarizes the attacks introduced in this section and lists the relevant standards where these attacks or the elements of mitigation strategies are described.

A. Data Breach

As a common privacy incident, a data breach is the disclosure of confidential or sensitive data in unauthorized access. This type of attack has a long history [46] in privacy and security challenges of any systems and is not limited to AI. Nevertheless, AI has increased the quality of the insight gained from big data and therefore, new vulnerabilities against data and privacy breaches have raised by AI. The data breach may happen in different phases of AI workflow [47]: Training, model, and inference phases. Confidentiality which is roughly an equal to privacy is the target of the adversary providing this attack.

As an early example of data breach attacks is re-identification where attackers used another dataset from the public electoral rolls of the city of Cambridge [35] to identify medical records. Additionally, a study on mobile phone metadata revealed that unique identification of 95% of individuals from a population of 1.5 million people, requires only 4 approximate location and time data points [32]. Different methods were implied to mask the sensitive information of individuals within the datasets [46]. Nonetheless, the evolution of big data and computational techniques such as AI systems provided new opportunities to violate data privacy in the process.

B. Bias in Data

The decisions achieved by AI systems can reinforce injustice and discrimination [38] in shortening candidates list for credit approval, recruitment, and criminal legal system [39]. Even though bias is not directly recognized as the privacy and security issue of big data, it is entangled with data and thereby can significantly impact the accuracy and accountability of the results. Among different types of bias [40] identified in AI systems, we focus on those which are correlated to data: i) *Sample bias* describing an unbalanced representation of samples in training data, ii) *Algorithm bias* which refers to the systematic errors in the system, and iii) *Prejudicial bias* indicates the incorrect attitude upon an individual data. Other types such as *measurement bias* that results from poorly measuring the outcome, are out of the scope of this paper. Bias is not a deliberate feature of AI systems, but rather the result of biases presented in the input data used to train the systems [48]. Hence, it targets the training phase and violates the *integrity* of an AI system.

Bias can target different attributes in decisions making including gender, race, age, national origin. In a project by MIT [25], known as Gender shade[2], the AI gender classification systems sold by giant technology companies (e.g., Microsoft, IBM, and Amazon) have been analyzed. The results

[2]http://gendershades.org/

TABLE II: Summary of the data privacy and security attacks in the AI workflow.

Attack	Security Goal (CIA)	Attack Examples	Developed / Under development Standards
Data Breach	Confidentiality	Re-identification [35] Risk of inference [36]	ISO/IEC CD 20547-4 [37], ISO/IEC PD TR 24028 [18]
Bias in Data	Integrity, Availability	Gender classification [25] Face recognition [38] Criminal legal system [39]	ISO/IEC NP TR 24027 [40], ISO/IEC PD TR 24028 [18]
Data Poisoning	Availability, Integrity	Self-driving car [27] Sentiment anlysis [41] Social media chatbot [42]	ISO/IEC PD TR 24028 [18]
Model Extraction	Confidentiality	Image recognition [28] Location data [43]	ISO/IEC PD TR 24028 [18]
Evasion	Integrity	Image classification [44] Spam emails [45] Self-driving car [11]	ISO/IEC PD TR 24028 [18]

of analysis in 2018, show a significant difference in the error rate of classifying darker-skinned female (up to 34.4%) in contrast to lighter-skinned males (0.8%). Some classification systems are considerably improved by 2019 [49] to reduce the error rate and yet the bias is not eliminated [50]. Bias is also found in a criminal legal system, Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) developed based on ML techniques to assess the sentencing and parole of convicted criminals. The purpose of COMPAS was to forecast the criminals who are most likely to re-offend [39]. However, the system has racial bias and tend to label black offenders almost twice higher risk than white offenders [51].

C. Data Poisoning

Data poisoning [26] is one of the most widespread attacks developed based on the idea of learning with polluted data. Its disruptive effects in industrial applications have attracted experts of the standard technical committee to investigate on the countermeasures and defence techniques [18]. The attack happens by injecting adversarial training data during the learning to corrupt the model or to force a system towards producing false results [52]. Therefore, the attack works in two ways: i) a common adversarial type is to alter the boundaries of the classifier such that the model becomes useless. This way the attacker aims the availability of the system. ii) the other type, however, targets the integrity of the system by generating a *backdoor* such that the attacker can abuse the system in his favor.

In a particular study on injecting poisoned samples to a deep learning model, it is shown that only 50 polluted samples are enough to achieve a 90% attack success rate in the system [53] while the accuracy remains almost the same. Early examples of data poisoning attacks are the worm signature generation [54], and spam filtering [55]. In another real world scenario of classifying the street signs in the U.S., a backdoor attack lead to the misclassification of the stop sign as the speed limit sign [27]. In social media, the data poisoning attack on Microsoft's chatbot, Tay, created a bot who made offensive and racist statements [42]. The bot was shut down only 16 hours after its launch. Sentiment analysis [41], malware clustering and detection [56]–[58] are the other target domains of this attack.

D. Model Extraction

The trained model is a valuable intellectual property in ML systems due to i) the big data source that is been used to train the model, and ii) the parameters (e.g., weights, coefficients) which generated for the model based on its function (e.g., classification) [18], [59]. The adversary's aim from the model extraction might be to infer record(s) that is used to train the model, thus, violates the confidentiality of the system. Based on how sensitive the trained data is (e.g., medical record), the attack can cause a significant privacy breach by disclosing sensitive information [29]. A reverse-engineering of ML model can happen by observing the input and output pairs [60] or by sending queries and analyzing the responses [59], where Tramer et al. prove that sending only hundreds of queries is sufficient enough to clone the same system with almost 100% accuracy.

Many ML techniques (e.g., logistic regression, linear classifier, support vector machine, and neural network) [61], [62] are shown to be vulnerable to this type of attack [59] and yet the proposed defense mechanisms are not sufficient enough to protect the privacy and security of data. In a study by Fredrikson et al. [9], [28], the authors report that having access to a face recognition model, they reproduce almost 80% of an individual's image from the training dataset. In a similar yet more successful attack on face recognition [63], attackers infer samples with a 100% success rate. Other examples of membership inference attack is also observed in location data disclosure [43], machine translation and video captioning [29], and medical diagnosis [63].

E. Evasion

Evasion is a popular common attack in which the attacker's aim is to evade detection by fooling the systems towards misclassification [31]. It happens in the apply phase of the AI workflow, where the real data is implied on the trained model. The well-known example of evasion attacks is the adversarial samples [18]. They are malicious samples that are designed adding a few chosen bytes to the original sample [30]. Even

though adversarial samples and poisoning data might look similar, they function differently. Considering a classifier, a data poisoning attack alters the classification boundary, however, adversarial samples modify the input samples to be classified in the wrong category. Hence, both lead to misclassification by targeting a different phase of AI workflow.

Adversarial samples are popular in comprising computer vision. In an experiment on autonomous vehicles [11], a couple of minor changes on the stop sign caused the learning model to misclassify the sign with a speed limit 45 sign. Even though for a human eyes such modifications does not affect the understanding of the street sign.

IV. Countermeasures and Privacy-preserving Solutions

This section describes an overview of the countermeasures and defense mechanisms of each particular attack mentioned in Section III.

A. Data Breach

The data protection and privacy techniques evolved during the time based on the growth of big data and the complexity of data analysis techniques. The purpose of these mechanisms is to ensure the confidentiality of data used for data analysis. Overall, the privacy-preserving techniques of big data can be categorized in three classes: Anonymization, De-identification, and Privacy-enhancing Techniques (PET). The privacy concerns of data is not a recent issue, started from data analysis on medial datasets in 1998, when the researcher find out that the anonymization is not sufficient solely to protect data privacy [46]. The sensitive data disclosure reports [32], [35] represent the deficiency of anonymization, where replacing clear identifier was enough solely to ensure the privacy and security of the data. Hence, the second level of mechanisms developed by k-*anonymity* [46] family, including l-*diversity* and t-*closeness* [17]. These techniques are suitable to mask sensitive information such as location-based data [64] to guarantee that the identity of records is not distinguishable in a dataset. The emergence of AI and ML techniques along with the increased complexity of big data, the conventional de-identification methods become obsolete [33]. Hence, PET was developed for privacy-preserving data analysis in various domains such as e-health [65], [66]. Fig. 2 describes these techniques according to the evolution of privacy-preserving techniques.

The next generation of the privacy-preserving approach is focused on the concept of *sending the code to the data*. The OPen ALgorithms (OPAL) project [67] has combined different mechanisms such as access-control protocols, aggregation schemes and develop a platform that allows third-parties (e.g., researchers) to submit algorithms that will be trained on data. The privacy of individuals, however, is guaranteed while data is being analyzed. Furthermore, Google's DeepMind has also developed a *verifiable data audit* which ensures that any interaction with health records data is recorded and accessible to mitigate the risk of foul play.

B. Bias in Data

To identify different types of bias several metrics are introduced in the literature [48] including difference in means, difference in residuals, equal opportunity, disparate impact, and normalized mutual information. Moreover, benefiting the metrics, approaches to mitigate AI bias are developed such as optimized prepossessing, reject option classification, learning fair representations, and adversarial debiasing [68]. Besides, a set of toolboxes are designed which are accumulated the identification metrics along with the mitigation approaches together as a framework for different ML algorithms. The purpose is to diagnose and remove AI biases if exists in the system. The available toolboxes are Lime, FairML [69], Google What-If and IBM Bias Assessment Toolkit [70] which is mostly used for face detection systems.

C. Data Poisoning

The feasibility of data poisoning attacks on ML algorithms such as Support Vector Machine (SVM) classifier is studied [71]. One common approach to detect the poisoned data is to identify the outlier (i.e., anomaly detection) since the injected data is expected to follow a different data distribution. Paudice et al. [72] developed their defense model against data poisoning based on anomaly detection. However, poisoned samples can evade anomaly detection if the adversary knows the data distribution. Hence, advanced techniques are required to defeat the attack. In [73], a method is proposed to perturb the incoming input and observe the randomness of the outcome. A low variance in the predicted classes represents malicious samples. Nelson et al. [74] proposed a technique to recognize and remove the poisoned data in the training dataset by separating the new joined input and calculate the accuracy of the model on them.

D. Model Extraction

Juuti et al. [75] proposed a method to detect model extraction attack by analyzing the distribution of consecutive API queries and compare it with benign behavior. One possible defense technique against model extraction is by training multiple models using different partitions of training data to each model. The techniques are proposed by Papernot et al. known as PATE [76]. Another approach to protect the learning model is to limit the information regarding the probability score of the model and degrade the success rate by misleading the adversary [77].

E. Evasion

Adversarial samples, as the most common evasion attacks, leas to misclassification only by small perturbations in the original inputs. Hence, a potential defense mechanism is to ensure that a small modification in the input cannot change the result significantly. Adversarial training is based on this technique to train the model based on the adversarial samples, however, with true labels such that it can avoid the noise [78]. In a similar approach by Deepfool [44] the ideas is to compute the perturbations which fool the classifier and thus quantify

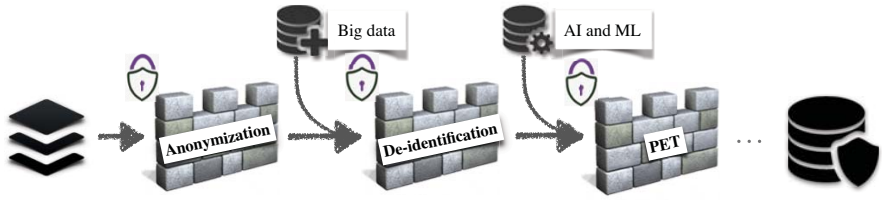

Fig. 2: A overview of the evolution of defense techniques for AI and big data analysis.

the robustness of the classifier. In another approach, the goal is to detect the adversarial samples from the original ones and therefore remove them from the dataset [79].

V. Summary

The huge volume, variety, and velocity of big data have empowered Machine Learning (ML) techniques and Artificial Intelligence (AI) systems. As privacy and security threats evolve, so too will the technology need to adapt – as well as the rules and regulations that govern the use of such technologies. The two perspectives of the research outcomes and standards development are considered in this study. We focus on challenges and threats of big data in the AI workflow by providing a review of the recent research literature, standard documents, and ongoing projects on this topic. Several projects are initiated by SDOs to investigate different aspects of big data privacy aspects and security issues. Even though most of the standards mentioned in this study are ongoing projects, they are expected to be published in the near future. One of the advantages standards can bring into research is a more coherent terminology, which is defined once and used later in subsequent projects. In contrast, researchers often use different terminologies for the same or similar concepts. Besides, according to the rapid growth of AI, developed road maps in standards can provide insights according to the demands and requirements of the market. Hence, it may provide opportunities for new research activities to address line with market needs.

Acknowledgement

This work is partially funded by the joint research programme University of Luxembourg/SnT–ILNAS on Digital Trust for Smart-ICT.

References

[1] D. Laney, "3d data management: Controlling data volume, velocity and variety," *META group research note*, vol. 6, no. 70, 2001.

[2] "Iso/iec tr 20547-2: Information technology – big data reference architecture – part 2: Use cases and derived requirements," International Organization for Standardization, Geneva, CH, Standard, 2018.

[3] K. Aditya and C. Wheelock, "Artificial intelligence market forecasts," Tractia, Tech. Rep., 2016.

[4] A. K. Keith Kirkpatrick, "Artificial intelligence use cases," Tractica, Tech. Rep., 2018.

[5] J. Chen, A. R. Kiremire, M. R. Brust, and V. V. Phoha, "Modeling online social network users' profile attribute disclosure behavior from a game theoretic perspective," *Computer Communications*, vol. 49, 2014.

[6] J. Chen, M. R. Brust, A. R. Kiremire, and V. V. Phoha, "Modeling privacy settings of an online social network from a game-theoretical perspective," in *IEEE Int. Conference on Collaborative Computing: Networking, Applications and Worksharing*, Oct 2013, pp. 213–220.

[7] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019.

[8] B. Biggio, G. Fumera, and F. Roli, "Security evaluation of pattern classifiers under attack," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 4, pp. 984–996, 2013.

[9] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. of ACM SIGSAC Conf. on Computer and Communications Security*, 2015.

[10] N. Carlini and D. Wagner, "Audio adversarial examples: Targeted attacks on speech-to-text," in *2018 IEEE Security and Privacy Workshops (SPW)*, 2018.

[11] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*.

[12] N. S. Labib, M. R. Brust, G. Danoy, and P. Bouvry, "Trustworthiness in IoT - a standards gap analysis on security, data protection and privacy," in *2019 IEEE Conference on Standards for Communications and Networking (CSCN'19)*, 2019.

[13] P. Cihon, "Standards for ai governance: International standards to enable global coordination in ai research & development," University of Oxford, Tech. Rep., 2019.

[14] (2019) Artificial intelligence, blockchain and distributed ledger technologies. CEN-CENELEC. [Online]. Available: https://www.cencenelec.eu/standards/Topics/ArtificialIntelligence/Pages\\/default.aspx

[15] H. Bae, J. Jang, D. Jung, H. Jang, H. Ha, and S. Yoon, "Security and privacy issues in deep learning," 2018.

[16] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. Leung, "A survey on security threats and defensive techniques of machine learning: A data driven view," *IEEE access*, vol. 6, 2018.

[17] N. Samir Labib, C. Liu, S. Dilmaghani, M. R. Brust, G. Danoy, and P. Bouvry, "White paper: Data protection and privacy in smart ict-scientific research and technical standardization," Tech. Rep., 2018.

[18] "Iso/iec pd tr 24028: Information technology – artificial intelligence (ai) – overview of trustworthiness in artificial intelligence," International Organization for Standardization, Geneva, CH, Standard.

[19] D. Poole, A. Mackworth, and R. Goebel, *Computational intelligence: a logical approach*. New York, U.S.: Oxford University Press, 1998.

[20] C. M. Bishop, *Pattern recognition and machine learning*. Springer Science, Business Media, 2006.

[21] "Iso/iec wd 22989: Artificial intelligence – concepts and terminology," International Organization for Standardization, Geneva, CH, Standard.

[22] "Iso/iec wd 23053: Framework for artificial intelligence (ai) systems using machine learning (ml)," International Organization for Standardization, Geneva, CH, Standard.

[23] S. Dilmaghani, M. R. Brust, A. Piyatumrong, G. Danoy, and P. Bouvry, "Link definition ameliorating community detection in collaboration networks," *Frontiers in Big Data*, vol. 2, 2019.

[24] A. M. Fiscarelli, M. R. Brust, G. Danoy, and P. Bouvry, "A memory-based label propagation algorithm for community detection," in *International Conference on Complex Networks and their Applications*. Springer, 2018, pp. 171–182.

[25] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Proc. of Conf. on fairness, accountability and transparency*, 2018.

[26] B. I. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, S.-h. Lau, S. Rao, N. Taft, and J. D. Tygar, "Antidote: understanding and defending against poisoning of anomaly detectors," in *Proc. of the 9th ACM SIGCOMM Conf. on Internet measurement*, 2009.

[27] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "Badnets: Evaluating backdooring attacks on deep neural networks," *IEEE Access*, 2019.

[28] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and b. r. U. S. S. y. . Ristenpart, Thomas", "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing."

[29] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *IEEE Symposium on Security and Privacy (SP)*, 2017, pp. 3–18.

[30] B. Kolosnjaji, A. Demontis, B. Biggio, D. Maiorca, G. Giacinto, C. Eckert, and F. Roli, "Adversarial malware binaries: Evading deep learning for malware detection in executables," in *26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 533–537.

[31] J. Zhang and X. Jiang, "Adversarial examples: Opportunities and challenges," *arXiv preprint arXiv:1809.04790*, 2018.

[32] Y.-A. De Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel, "Unique in the crowd: The privacy bounds of human mobility," *Scientific reports*, vol. 3, 2013.

[33] Y.-A. de Montjoye *et al.*, "Response to comment on "unique in the shopping mall: On the reidentifiability of credit card metadata"," *Science*, vol. 351, 2016.

[34] J. Andress, *The basics of information security: understanding the fundamentals of InfoSec in theory and practice*. Syngress, 2014.

[35] L. Sweeney, "Simple demographics often identify people uniquely," *Health (San Francisco)*, vol. 671, 2000.

[36] E. Jahani, P. Sundsøy, J. Bjelland, L. Bengtsson, Y.-A. de Montjoye *et al.*, "Improving official statistics in emerging markets using machine learning and mobile phone data," *EPJ Data Science*, vol. 6, 2017.

[37] "Iso/iec cd 20547-4: Information technology – big data reference architecture – part 4: Security and privacy," International Organization for Standardization, Geneva, CH, Standard.

[38] M. Wall. (2019, Jul.) Biased and wrong? facial recognition tech in the dock. BBC. [Online]. Available: https://www.bbc.com/news/business-48842750

[39] S. X. Zhang, R. E. Roberts, and D. Farabee, "An analysis of prisoner reentry and parole risk using compas and traditional criminal history measures," *Crime & Delinquency*, vol. 60, 2014.

[40] "Iso/iec np tr 24027: Information technology – artificial intelligence (ai) – bias in ai systems and ai aided decision making," International Organization for Standardization, Geneva, CH, Standard.

[41] A. Newell, R. Potharaju, L. Xiang, and C. Nita-Rotaru, "On the practicality of integrity attacks on document-level sentiment analysis," in *Proc. of the Artificial Intelligent and Security Workshop*, 2014.

[42] J. Vincent. (2016) Twitter taught microsoft's ai chatbot to be a racist asshole in less than a day. [Online]. Available: https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[43] A. Pyrgelis, C. Troncoso, and E. D. Cristofaro, "Knock knock, who's there? membership inference on aggregate location data," *CoRR*, 2017.

[44] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proc. of the IEEE Conf. on computer vision and pattern recognition*, 2016.

[45] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, 2018.

[46] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," Technical Report, SRI International, Tech. Rep., 1998.

[47] M. Al-Rubaie and J. M. Chang, "Privacy-preserving machine learning: Threats and solutions," *IEEE Security & Privacy*, vol. 17, 2019.

[48] J. H. Hinnefeld, P. Cooman, N. Mammo, and R. Deese, "Evaluating fairness metrics in the presence of dataset bias," *arXiv preprint arXiv:1809.09245*, 2018.

[49] I. D. Raji and J. Buolamwini, "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products," in *AAAI/ACM Conf. on AI Ethics and Society*, 2019.

[50] A. Amini, A. Soleimany, W. Schwarting, S. Bhatia, and D. Rus, "Uncovering and mitigating algorithmic bias through learned latent structure," 2019.

[51] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How we analyzed the compas recidivism algorithm," *ProPublica*, 2016.

[52] J. Steinhardt, P. W. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Advances in neural information processing systems*, 2017.

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

[54] J. Newsome, B. Karp, and D. Song, "Paragraph: Thwarting signature learning by training maliciously," in *International Workshop on Recent Advances in Intrusion Detection*, 2006.

[55] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. Rubinstein, U. Saini, C. A. Sutton, J. D. Tygar, and K. Xia, "Exploiting machine learning to subvert your spam filter." *LEET*, vol. 8, 2008.

[56] B. Biggio, K. Rieck, D. Ariu, C. Wressnegger, I. Corona, G. Giacinto, and F. Roli, "Poisoning behavioral malware clustering," in *Proc. of the workshop on artificial intelligent and security*, 2014.

[57] S. Misra, M. Tan, M. Rezazad, M. R. Brust, and N.-M. Cheung, "Early detection of crossfire attacks using deep learning," *arXiv preprint arXiv:1801.00235*, 2017.

[58] M. Rezazad, M. R. Brust, M. Akbari, P. Bouvry, and N.-M. Cheung, "Detecting target-area link-flooding ddos attacks using traffic analysis and supervised learning," in *Future of Information and Communication Conference*. Springer, 2018, pp. 180–202.

[59] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction apis," in *25th USENIX Security Symposium*, 2016.

[60] S. J. Oh, M. Augustin, B. Schiele, and M. Fritz, "Towards reverse-engineering black-box neural networks," *arXiv preprint arXiv:1711.01768*, 2017.

[61] B. Wang and N. Z. Gong, "Stealing hyperparameters in machine learning," in *IEEE Symposium on Security and Privacy (SP)*, 2018.

[62] D. Lowd and C. Meek, "Adversarial learning," in *Proc. Int. Conf. of the 11th ACM SIGKDD*, ser. KDD '05, 2005.

[63] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "Logan: Membership inference attacks against generative models," 2019.

[64] F.-J. Wu, M. R. Brust, Y.-A. Chen, and T. Luo, "The privacy exposure problem in mobile location-based services," in *IEEE Conf. on Global Communications (GLOBECOM)*, 2016.

[65] F. K. Dankar and K. El Emam, "The application of differential privacy to health data," in *Proc. of the Joint EDBT/ICDT Workshops*, 2012.

[66] S. Dilmaghani, "A privacy-preserving solution for storage and processing of personal health records against brute-force attacks," Master's thesis, University of Bilkent, Turkey, 2017.

[67] OPAL. (2017) Open algorithms. [Online]. Available: http://www.opalproject.org/

[68] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *IEEE 12th Inter. Conf. on Data Mining*, 2012.

[69] J. A. Adebayo, "Fairml: Toolbox for diagnosing bias in predictive modeling," Ph.D. dissertation, MIT, 2016.

[70] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic *et al.*, "Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv preprint arXiv:1810.01943*, 2018.

[71] B. Biggio, I. Corona, B. Nelson, B. I. Rubinstein, D. Maiorca, G. Fumera, G. Giacinto, and F. Roli, "Security evaluation of support vector machines in adversarial environments," in *Support Vector Machines Applications*. Springer, 2014.

[72] A. Paudice, L. Muñoz-González, A. Gyorgy, and E. C. Lupu, "Detection of adversarial training examples in poisoning attacks through anomaly detection," *arXiv preprint arXiv:1802.03041*, 2018.

[73] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "Strip: A defence against trojan attacks on deep neural networks," *arXiv preprint arXiv:1902.06531*, 2019.

[74] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. Rubinstein, U. Saini, C. Sutton, J. Tygar, and K. Xia, "Misleading learners: Co-opting your spam filter," in *Machine learning in cyber trust*. Springer, 2009.

[75] M. Juuti, S. Szyller, S. Marchal, and N. Asokan, "Prada: protecting against dnn model stealing attacks," in *IEEE European Symposium on Security and Privacy (EuroS&P)*, 2019.

[76] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable private learning with pate," *arXiv preprint arXiv:1802.08908*, 2018.

[77] T. Lee, B. Edwards, I. Molloy, and D. Su, "Defending against machine learning model stealing attacks using deceptive perturbations," *arXiv preprint arXiv:1806.00054*, 2018.

[78] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[79] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," in *Proc. of the ACM SIGSAC Conf. on Computer and Communications Security*, 2017.