UC Irvine
UC Irvine Previously Published Works

Title

Trustworthy AI Inference Systems: An Industry Research View

Permalink

https://escholarship.org/uc/item/5rj2w6nf

Authors

Cammarota, Rosario
Schunter, Matthias
Rajan, Anand
et al.

Publication Date

2020-08-10

Copyright Information

Peer reviewed

Trustworthy AI Inference Systems: An Industry Research View

ROSARIO CAMMAROTA, MATTHIAS SCHUNTER, and ANAND RAJAN, Intel Labs, USA

FABIAN BOEMER, Intel AI, USA

AMOS TREIBER, THOMAS SCHNEIDER, EMMANUEL STAPF, and AHMAD-REZA SADEGHI, TU Darmstadt, Germany

DANIEL DEMMLER and JOSHUA STOCK, University of Hamburg, Germany

CHRISTIAN WEINERT, Royal Holloway, University of London, UK

ÁGNES KISS, CISPA Helmholtz Center for Information Security, Germany

HUILI CHEN, SIAM UMAR HUSSAIN, SADEGH RIAZI, FARINAZ KOUSHANFAR, SARANSH GUPTA, TAJANA ROSING, and KAMALIKA CHAUDHURI, UC San Diego, USA

HAMID NEJATOLLAHI, NIKIL DUTT, and MOHSEN IMANI, UC Irvine, USA

AYDIN AYSU and ANUJ DUBEY, North Carolina State University, USA

KIM LAINE, Microsoft Research, USA

FATEME SADAT HOSSEINI and CHENGMO YANG, University of Delaware, USA

ERIC WALLACE and PAMELA NORTON, Borsetta Research, USA

In this work, we provide an industry research view for approaching the design, deployment, and operation of trustworthy Artificial Intelligence (AI) inference systems. Such systems provide customers with timely, informed, and customized inferences to aid their decision, while at the same time utilizing appropriate security protection mechanisms for AI models. Additionally, such systems should also use Privacy-Enhancing Technologies (PETs) to protect customers' data at any time.

To approach the subject, we start by introducing current trends in AI inference systems. We continue by elaborating on the relationship between Intellectual Property (IP) and private data protection in such systems. Regarding the protection mechanisms, we survey the security and privacy building blocks instrumental in designing, building, deploying, and operating private AI inference systems. For example, we highlight opportunities and challenges in AI systems using trusted execution environments combined with more recent advances in cryptographic techniques to protect data in use. Finally, we outline areas of further development that require the global collective attention of industry, academia, and government researchers to sustain the operation of trustworthy AI inference systems.

1 INTRODUCTION

At a macroscopic level, the players in the current AI and privacy landscape are customers (including end-users), institutions, and regulators. Customers of digital systems continue maturing awareness of the risks associated with the mishandling of and the malicious access to their private information. Such risks include loss of competitiveness, financial loss, brand erosion, identity theft, and more profoundly human rights violations. Moreover, the expectation for AI inference systems to become integral parts of modern digital systems and the lack of understanding of how some AI models operate further exacerbate customers' concerns. Institutions aim to capitalize on building more accurate AI inference systems from shared data belonging to multiple stakeholders for providing better services to their customers, including end users [6]. The leading technology companies improve the effectiveness of their platform recommendation systems by embedding their customers' preference patterns in sophisticated AI models [2]. Such models are part of inference systems to deploy and operate in the cloud, in the core network, on devices, or a mixture

thereof. Often, the raw data processed by such systems can carry private information, such as gender, location or political views, and hence *requires privacy protection at any time* [4]. Simultaneously, the AI model and its coefficients (e.g., weights and biases) often require IP protection. Institutions are also aware of the regulatory and legal risks associated with mishandling private data and data breaches [42, 57], rendering cyber insurance policies indispensable for more and more companies [126]. Regulators create and enact regulations to protect the rights of customers and customers of customers. The enforcement of such privacy-related regulations (e.g., the European GDPR [4] in the EU, CCPA [5], and the HIPAA [3] in the US) have already resulted in substantial fines to institutions starting in 2019[1]. In addition, the European Commission has proposed a regulatory framework [55] which explicitly mentions trustworthiness as one of the criteria AI systems should fulfill.

At a microscopic level, trends in AI and the crucial dimension of privacy influence the way researchers and technologists think about privacy technologies in the context of the complex interaction between customers, institutions, and regulators. Designing technologies to protect data confidentiality at any time in the data lifecycle is not a new concept. In the 1990s, Ann Cavoukian introduced the concept and principles of *privacy by design* that requires end-to-end data protection, which includes the protection of data in use (including the input, the intermediate data, and the output of a computation) [72]. Privacy-enhancing cryptographic techniques, such as secure multiparty computation or homomorphic encryption, can protect data in use in compliance with the data privacy regulations (e.g., GDPR). However, such methods are not yet widely adopted mainly due to their programming complexity and computational overhead compared to insecure native applications. The renewed awareness of customers' risks with data digitization and AI, and the expectation of AI to play in the global economy in the foreseeable future [10], have accelerated industry and academic researchers' efforts to advance privacy-enhancing cryptographic techniques towards the realization of broader market adoption and application [7].

There resides complex infrastructures in-between the micro- and macroscopic levels, comprising the cloud, the end devices, and base stations, that abstract computation, storage, and communication resources. Such infrastructures represent the contact point between customers, institutions, regulations, and AI systems. The infrastructures are also where a sustainable deployment of trustworthy AI inference systems is needed to address the mainstream concerns associated with data privacy and IP protection.

Deployment and Operation. Deploying and operating trustworthy AI inference systems is challenging. In its most straightforward instance, an AI inference system captures the interactions among three players: (*i*) the customer, who owns the input data to an AI service; (*ii*) the service provider, who owns, deploys, and operates the AI inference service; and (*iii*) the infrastructure provider (e.g., a cloud provider or a network operator).

Such parties are mutually distrustful, can collude, and can be malicious while aiming for different goals. The AI inference service, often an IP owned (either created or acquired) by the service provider, needs to be protected from the infrastructure provider and the customer. The customer's input data may carry privacy-sensitive information (e.g., in the case of medical records), and requires protection from both the service and infrastructure provider. Furthermore, if the provider experiences security breaches, the input data, the intermediate results, and the computation's output should remain confidential. Finally, the infrastructure provider requires protection from both the service provider and the customer. These aspects all require careful consideration when deploying solutions.

---

[1] As a consequence of data breaches, British Airways PLC received a fine of GBP 183M [42], Facebook received a fine of USD 5B [57], to name a few.

Security Mechanisms (TEE). Approaching the challenge of deploying and operating trustworthy AI systems in its entirety requires thinking beyond traditional security mechanisms for confidentiality, integrity, and IP protection such as a Trusted Execution Environment (TEE) [19, 26, 123]. TEEs allow protecting the service provider's data at rest with encrypted storage, input data in transit with secure channels, and the software, the intermediate data, and outputs from the infrastructure provider. However, the input, intermediate, and output data (data in use) appear in the clear in the TEE. Protection of data in use should be seen at its core through the lens of the social relationship governing disclosure between and among the customer and the service provider via the platforms that collect, analyze, and manipulate information for some purpose, without necessarily requiring parties to trust each other fundamentally. Occasionally, offering TEE protection to customers' data can be sufficient, but it is not sufficient in the general case. It is paramount to provide service providers' access to customers' data for processing purposes, while not revealing the data to the processing party.

Advanced Cryptographic Techniques (MPC and HE). To provide service providers access to customers' data, while not revealing the data to the processing party, and to protect customers' data even in the presence of data breaches at the service or infrastructure provider, technologists can resort to advanced cryptographic techniques. For example, secure Multi-Party Computation (MPC) and Homomorphic Encryption (HE) enable computation on private data without ever having plaintext access. The absence of having the corresponding decryption information enhances customers' data protection even in the case of data breaches. Fortunately, there has been an explosion of innovation in advanced cryptographic techniques for computing on encrypted data in recent years—such innovation promise to lower technology barriers currently hindering corporations in adopting the said technologies within existing workflows. Proofs-of-concept have appeared in the context of AI inference, specifically in Deep Learning (DL) [16, 21, 23, 39, 61, 77, 85, 106].

Hybrid Solutions. Hence, while it is foreseeable for such techniques to percolate into technology, technologists must bear in mind that the development, deployment, and operation of an AI inference system must protect customers' data at any time, as well as the security of the AI service and its infrastructure. Technology providers should foresee a future in which the technology deploys symbiotic combinations of current system security practices and advanced cryptographic techniques to build secure, private, and trustworthy AI inference systems.

2 TAIL AND HEADWINDS

In this section, we highlight catalysts and design challenges for the real-world deployment of trustworthy AI inference systems. The discussion articulates the elements of an AI inference system, roughly organized in the categories application, software, infrastructure, and hardware.

Application. In the application category, vision and specifically face recognition technologies are mainstream. In the private sector, one of the most impactful events for business innovation is the approval and adoption of AI for medical and diagnostic applications (e.g., the FDA green-lighting AI-as-a-medical-device [58]). Additionally, early commercial applications are taking off in security, retail, and consumer electronics, in which face recognition is quickly becoming a dominant form of biometric authentication. In the public sector, however, there is strong opposition for the deployment of face recognition technologies for surveillance-related applications, as indicated by the slowdown of the US Department of Homeland Security projects [129]. Even Amazon, Microsoft, and IBM recently decided to limit the commercialization of their face recognition technologies due to privacy and other societal concerns [20, 29]. Though it is clear that the ability to compute on

encrypted data should be technically sufficient to address the privacy concerns, meaningful adoption of cryptographic privacy technologies needs other developments than just technology. *Other humanity disciplines* are needed to understand the interaction with humans. In particular, both customers and institutions must have a thorough understanding of privacy technologies via educational paths, and agreements on the use of such technologies via the establishment of standards and best practices. Relevant new challenges have recently been identified [9].

Software. In the software level category, the barrier to entry for the AI space is lower than ever, thanks to open-source software. Following TensorFlow and Keras, major big-tech players have made available various front-end frameworks for developers to choose from, including Apache® MXNet. Additionally, middle-end and back-end optimizing frameworks (e.g., Apache® TVM, Intel™ nGraph, OpenVino®, and PlaidML [115]) were developed to facilitate the optimization of graph computation and their mapping on a variety of hardware targets such as CPU, GPU, and FPGA. The software level exhibits the path of least resistance in percolating privacy technologies in AI systems, as illustrated by the appearance of proofs-of-concepts in the literature in the past few years [21, 23, 44, 85].

Infrastructure. In the infrastructure category, the need for real-time decision making is pushing AI closer to the edge, from the core network to demarcation points and down to end devices. This trend gives devices the ability to process information locally and respond faster. Furthermore, the data remains in proximity or mostly close to the data owner, which inherently helps with privacy. Even so, in the presence of stolen devices and data breaches, it would be desirable to lift higher the bar of data confidentiality with privacy-enhancing cryptographic techniques.

Hardware. In the hardware category, many players in the semiconductor industry and startups are focused on building chips exclusively for AI workloads. It is noteworthy that semiconductor players use specialized hardware and native data type specialization to fit in hardware constraints and reach adequate performance by trading efficiency, accuracy, and speed. However, the sway in the trade-off is domain-specific and application-dependent. For example, in the case of face recognition to unlock devices, the computation needs to fit within tenths of a millisecond. Accuracy is essential, but a delay in response hinders usability. In other cases, such as medical diagnosis, speed is relatively not critical, but a lack of accuracy can have fatal consequences for the customers, leading to a loss of customers' trust. Similarly, both high-accuracy and timeliness are required in predicting an evacuation window given an expanding fire.

Hardware-assisted technologies such as TEE have existed for a long time to protect proprietary software IPs. Solutions exist to protect AI models from the infrastructure hosting the accelerator [60]. Introducing hardware aids for advanced cryptographic techniques remains challenging. Such techniques often rely on data types for which mainstream devices are not designed for (e.g., high-degree polynomials with large coefficients in the case of homomorphic encryption). There is also a lack of standardization of TEE architectures making programming difficult. The design of new standardized hardware for privacy-enhancing cryptographic techniques would be a game changer [71, 100, 103]. Making such techniques inexpensive would create opportunities inconceivable with current software implementations optimized for existing hardware targets.

3   PRIVACY TECHNOLOGIES PENETRATION

Including privacy technologies into different facets of AI system design via security mechanisms and advanced cryptographic techniques has been a very active field of research in recent years. Most research focuses on machine learning (ML) tasks, providing accurate privacy-preserving

classification using artificial, specifically convolutional neural networks. In Machine Learning-as-a-Service (MLaaS), offered by all major cloud service providers, privacy technologies can allow clients to issue classification queries without revealing the potentially sensitive information that should be classified in the clear. In some cases, both model coefficients (weights and biases) and the functional form of the service providers' model are hidden from clients. They represent intellectual property (IP) and might contain traces of sensitive training data. In other cases, only the query to the model is protected.

TEE. Trusted Execution Environment (TEE) architectures such as Intel® Software Guard eXtensions (SGX), ARM® TrustZone®, and Sanctuary [27], to name a few, are hardware-assisted security architectures that create a level of security and trust that goes beyond the protection capabilities of commodity Operating Systems (OS).

Solutions that use TEEs alone [19, 26, 123] can protect the IP and are often the most efficient since many TEE implementations work at the native speed of the CPU. However, the software within the TEE processes the input data (or query), intermediate results, and outputs in the clear. There exist practical cases where protection via a TEE cannot be enforced. In such cases, other techniques such as proof of ownership [110] can be used to provide evidence that a service has been counterfeited and that illegitimate uses of the service occurred.

MPC. Secure multi-party computation (MPC) protocols [62, 134] are cryptographic protocols that allow multiple parties to jointly compute a publicly known function on their private inputs while revealing no information other than the result of the computation. Therefore, MPC emulates a trusted third party's behavior within a provably secure cryptographic protocol without the need to rely on such a trusted third party.

In solutions that use MPC protocols alone (e.g., [16, 85, 106]), the participants learn the architecture of the trained AI model. MPC solutions protect both the query to the model and the model coefficients, but not the functional form of the model. Furthermore, when the model owner does not own the computational infrastructure (e.g., the model owner serves the model via a cloud provider), the service deployment must resort to a trusted execution environment (TEE). The TEE protects the model coefficients and allows the model owner to outsource the model coefficients to multiple non-colluding parties that run MPC [78].

Hiding even the architecture of the AI model with MPC techniques is possible by using so-called universal circuits that can be programmed by the service provider to emulate any function up to a given size [12, 64, 81, 83, 89, 112, 125, 135]. Securely evaluating such a universal circuit with an MPC protocol results in private function evaluation (PFE), where the function itself is kept private. However, universal circuits have an inevitable logarithmic overhead [125] and therefore are not suitable for huge functions that occur in machine learning applications.

HE. Homomorphic Encryption (HE) is a potent tool to preserve the privacy of data and users. An encryption scheme that has homomorphic properties enables computations on encrypted data. Anyone can perform computations over a ciphertext without the need to perform decryption. In practice, only limited computation on encrypted data is feasible today, creating a challenge for finding AI model architectures that are compatible with these limitations.

HE could allow IP protection inherently when using Fully Homomorphic Encryption (FHE) constructions [24, 38, 52, 56, 59]. Due to the bootstrapping procedure to refresh ciphertexts noise, the parameter selection to instantiate an FHE schema can become independent from the complexity of the function to evaluate. Hence, only the party performing the inference learns (knows) the model. If such a party is the cloud provider (e.g., AI serviced via Google's or Microsoft's infrastructure), the model can be protected with the security schemes used to protect the cloud infrastructure and

its services. If the party owning the model serves the model via a cloud infrastructure, the deployment can utilize a TEE to protect the model from the cloud infrastructure. In either case, fully homomorphic encryption methods can evaluate the model in the clear on encrypted queries while allowing computation on encrypted data owned by two (or more) distrusted parties. Industry has recently picked up on implementations and performance evaluation in this field and commercial applications are to be expected [13, 39, 46].

DP. Differential privacy (DP) [54] is a statistical notion of privacy where the goal is to protect the privacy of individuals in the training data from an adversary who sees the trained model. Privacy is guaranteed by ensuring that the participation of a single person in the training data does not change the probability of any outcome by much. This implies that an adversary who sees the output of a differentially private algorithm cannot make any inferences about a person in the training data with high confidence that they could not make if this person had not been in the training data at all.

In machine learning, differential privacy is mostly provided by adding noise during the training process. Noise can be added in several ways – to the data itself, to a classifier built on the data [35, 111], or to a loss function or objective [35]. Currently the most common method for differentially private machine learning is to add noise at each iteration of a stochastic gradient descent process during training [8, 17, 116]. This provides good privacy, but the added noise results in a loss of statistical efficiency – measured by model quality per sample size. Model quality can be measured by classification accuracy (for classification models) or test log-likelihood (for generative models), and usually the loss of quality due to privacy is lower when a large amount of training data is present. While much progress has been made in the past years, ensuring high statistical efficiency remains a central research problem in differentially private machine learning.

Hybrid Solutions. In addition to the protection profile, there are performance considerations that call for hybrid solutions that realize privacy-, security-, and ultimately trustworthiness-preserving building blocks. A straightforward implementation of such AI tasks with one advanced cryptographic technique only will most likely not result in practical solutions in terms of latency, throughput, and costs for computation power or network traffic in real-world applications. For example, in state-of-the art instantiations of Yao's garbled circuits MPC protocol [109], 197 bits of data must be sent for each *binary* AND gate which is expensive for large multiplication circuits.

Hence, alternative hybrid approaches may be required to make privacy preserving MLaaS practically viable. Examples are the development of more efficient sub-protocols for specific tasks (e.g., matrix multiplication), data and network pre-processing, combinations of MPC protocols with HE or TEEs, optimized mixing of different MPC protocols, and targeting more realistic and practical security models and security levels.

Modern MPC protocols are a noteworthy example of hybrid protocols. Today's most efficient MPC protocols are hybrid protocols that combine different MPC protocols and potentially also HE for different sub-tasks [15, 16, 21, 47, 66, 96, 106]. There is also compiler support for automation [31, 34, 75]. This results in highly efficient protocols for private AI (e.g., for deep neural networks) [16, 21, 77, 90, 95, 97, 104–106, 110], for decision trees [15, 16, 80], and for new machine learning models such as Hyperdimensional computing [74] or sum-product networks [124]. However, most of the prior art provides solutions that are not yet easy-to-use by data scientists and have mostly been demonstrated on simple deep neural networks and small to medium datasets (e.g., MNIST [87] or CIFAR-10 [84]). In comparison to advanced cryptographic techniques, TEE architectures provide several orders of magnitude better performance when used for protecting AI systems [60, 123]. Furthermore, TEE architectures can be used to alleviate overheads introduced by advanced cryptographic techniques. For example, in [127], a TEE, specifically Intel® SGX, was

used to obviate the severe overhead in the software when performing bootstrapping in FHE. The usage of a TEE architecture always requires some trust in the TEE provider. However, it is worth noting that trusting the TEE provider is a practice that is well established in the semiconductor industry to protect third party IP [32].

4 ECOSYSTEM DEVELOPMENT

Although there have been significant advances from the ongoing effort of academic, industry, and government research groups, further research and development efforts are needed to realize a meaningful commercial adoption of trustworthy AI inference systems. For example, none of the privacy-preserving techniques (building blocks) presented above are easy to program, not to mention proving the security of finalized systems.

Compilers. For MPC and HE, recently automation tools, specific compilers, and optimizers have been designed that help developers build correct and performing solutions with MPC/HE libraries. For MPC, such tools have been available for a while [28, 47, 66, 94], including projects where tools designed for hardware development have been re-purposed to optimize according to the cost models for encrypted "gates" [48, 117]. For example, CHET [45] is a framework optimized explicitly for neural network inference. EVA [44] builds on CHET and adds a general-purpose compiler for a specific HE scheme. Other projects [37, 44, 45, 96] provide frameworks for targeting multiple library backends, providing a high-level unified Intermediate Representation (IR) or Application Programming Interface (API). While promising, these tools are far from being mature, and mostly target advanced cryptographic techniques in the same building block family (e.g., MPC or HE).

Complex applications. Mapping complex applications onto advanced cryptographic techniques is still an ad hoc process and requires almost always cryptographic expertise to be done correctly. In the case of DL workloads, much of the literature shows that solutions with advanced cryptographic techniques can be handcrafted, but only recent work has started to show the execution of full DL workloads. Frameworks such as Intel® nGraph-HE [21–23], TF-Encrypted [43], and CrypT-Flow [85] allow seamless execution of pre-built neural networks on HE and MPC building blocks. CryptoSPN [124] uses sum-product networks (SPNs) for private inference and is integrated with an ML toolchain. The literature has focused on cases in which the service provider is also the owner of the infrastructure. In this case, the data owner also trains and owns the model. As new business models shift towards decentralizing the roles of the various owners in the lifecycle of AI systems, designing more complex system architectures becomes necessary, specifically combining TEEs and advanced cryptographic techniques. By themselves, most TEEs are also not easy to program. Fortunately, research programs such as Graphene [63], a joint effort between Intel and academic partners, provides a solution to program SGX enclaves seamlessly and securely.

5 RESEARCH AND STANDARDS

The trend of privacy research and technology development will continue to mature current technologies ready for mass deployment. We consider hybrid solutions as viable to build trustworthy AI systems from a technology standpoint. Such solutions open opportunities to sustain the required system performance while preserving security and privacy properties. With the progress made so far, there are many areas of opportunities for further innovation developments and addressing challenges for reaching meaningful adoption.

Decomposition. At a fundamental level, methods appearing in the literature are tied to specific ML workloads, typically DL prototype networks. Thus, there is a lack of a broad understanding of how to decompose an ML workload onto a set of cryptographic mechanisms for security, data

privacy, performance, and energy efficiency. Filling this knowledge gap becomes essential because new ideas and advances in ML are being proposed and adopted regularly, and the fact that privacy-preserving building blocks are aggressively introduced and refined.

Integration. There is a need and opportunity to increase the integration of privacy-preserving building blocks in established ML development tools such as ML frameworks and graph compilers. For example, Intel® nGraph-HE-ABY [21] is an industry-class and open-source framework supporting both HE and MPC. Beyond the integration of building blocks, such a (development/testing) framework should also include automatic mechanisms to map ML workloads onto combinations of privacy-preserving building blocks for data privacy (e.g., HE versus MPC variants) and performance objectives. Additionally, it should be easy to incorporate new mechanisms as advances in privacy technologies appear in the literature to easily repair or improve existing systems.

Deployment. There is also a need and opportunity to develop run-time frameworks that can reason about the available resources of computation, storage, communication, and security requirements, to automatically and expediently orchestrate the workload component for efficiency. This is particularly important due to the additional resource demand imposed with the adoption of modern privacy-preserving mechanisms and their parameter selection options.

Benchmarks. A fundamental lack of benchmarks leaves another gap to be filled as many experiments restrict themselves to different tasks and assumptions. Systematic comparisons like, e.g., [65] for a critical mass of experiments across AI and other tasks are required at a sufficient scale to draw conclusions that can drive technology and best practices development, transfer and adoption.

Hardware. New hardware is also an important topic for future research. It is paramount to design a programmable processor that provides end-to-end data security and privacy. Although new processor technology has evolved to serve complex encryption tasks more efficiently, data movement costs between the processor and memory still hinder the higher efficiency of applications relying on advanced cryptographic techniques for privacy protection [73, 113]. To address this issue, active research in industry and academia need to focus on designing novel architectures, including but not limited to non-von Neumann compute architectures such as in-memory or near-data computing [14, 100].

Standardization. Finally, there is a need for the development of global standards and best practices such as definitions, technical foundations and application standards to facilitate the broad deployment of advanced cryptographic mechanisms [11, 30]. International standards work under ISO/IEC JTC 1 Information Security includes privacy technologies for trustworthy AI systems [98, 101], and early-stage standardization on MPC and FHE techniques.

Provenance. Integrity and security of AI systems are required for validating privacy requirements. Immutable ledgers for data authenticity, trusted electronic supply chains, and verified neural networks for AI inference systems provide necessary trust to enable wide distribution of the technology. There is a need for active research in integrating technologies like blockchain and distributed ledger technology into AI systems to ensure provenance and integrity throughout the manufacturing and lifecycle of the AI system components and to ensure privacy preserving requirements are met [119, 132].

Fault Resiliency. AI models are vulnerable to various types of faults, such as permanent stuck-at defects, random bit-flip, and thermal noise [82, 121], which makes reliability critical issues in AI hardware. Both device aging and malicious activities can contribute to such reliability issues.

Circuit vulnerabilities can be leveraged by attackers to intentionally cause the hardware to fail under a variety of attack vectors, including row hammer, bit-flip, gradient descent, and backdoor attacks [40, 41, 67, 92, 102, 118]. Although recent work has investigated errors in AI models [36, 70, 88, 130, 131], they mostly focus on tolerating manufacturing defects and overlook errors caused by device aging or malicious intent. Such solutions also require non-trivial retraining and data/hardware redundancy, limiting their applicability to edge devices with tight hardware budget and limited training capability. Hence, to monitor and maintain reliability and trustworthiness of AI accelerators, it is desirable to develop self-test and self-healing techniques that integrate test, diagnosis, and recovery loops into the system. Such techniques should target both hard defects and soft errors. They should perform a non-destructive self-test periodically to detect errors and pinpoint defective cells if any. Upon detecting any error-induced accuracy loss, such techniques should offer a retraining-free self-healing process to rescue the AI model's accuracy. The interaction of such techniques and privacy technologies is an open topic of research.

Side-Channel Resiliency. Although black-box model extraction techniques [76, 122] are ever-evolving by adopting the cryptanalytic methods [33], they are still economically infeasible, since they need millions of queries to steal a model with high fidelity. Side-channel analysis on inference systems [18, 49–51, 53, 68, 69, 91, 99, 120, 128, 133], by contrast, can succeed with significantly fewer tests and are harder to prevent as we learned from the research on cryptographic engineering. Side-channel analysis can steal both the input data of the customer [128], as well as the model of the service provider [18], the general architecture of the deployed AI system [91], and the detailed, bit-level values of its coefficients [51]. The defenses against digital side-channels, such as timing side-channels, are relatively more natural to establish for specialized accelerators (as opposed to general-purpose engines) because the design tools enable cycle-accurate control and simulation. Building defenses to prevent physical side-channel leakages such as via power consumption or electromagnetic radiation is challenging and needs tuning for the specifics of the target algorithm and its underlying implementation. Research has shown the first proofs-of-concept by extending the techniques from cryptographic engineering and empirically validating the physical side-channel security [50, 51]. Further research is needed to achieve provably-secure, composable, and low-cost solutions that can be deployed in real-world applications of trustworthy AI systems. However, such research is unlikely to occur without significant investment, given that it is still a challenge for cryptographic systems after two decades of work. One way to accelerate the process is government involvement in standardizing the defenses [25].

6 PRIVACY AND EXPLAINABILITY

The performance of AI models depends on the quality of the training data. Training datasets with inherent bias lead to models replicating that bias, sometimes resulting in positive feedback loops where biased decisions are made based on decisions from biased models, creating more biased data that the models are further trained on. Many ML models, such as DNNs, are poorly explainable (i.e., it is hard to explain post-hoc what contributed to a specific incorrect or biased prediction). It can be hard to understand how well or how poorly the model will behave on unexpected inputs. Privacy technologies can further complicate this situation: if the model owners cannot see the query or the result, they have little hope of detecting incorrect or biased predictions, possibly thwarting good intentions to produce fair and unbiased ML applications. Novel strategies, such as Hyperdimensional computing, show promise in supporting secure and private computation while being explainable and easily updateable online [74, 79]. For a broader spectrum of applications, many explainable AI (XAI) tools have been developed in recent years. They can be helpful for explaining single decisions of arbitrary black-box classifiers [93, 107, 108] and for visualizing the

general influence of single features [1]. In addition, there are many domain-specific XAI tools, e.g., saliency maps for image classification [86, 114].

7 CONCLUSION

In this work, we identify vehicles to percolate privacy technologies for AI inference, such as open-source frameworks, infrastructure, and hardware design. We foresee architectural solutions, including hybrid methods with security mechanisms and advanced cryptographic techniques as viable innovations for designing, building, deploying and operating secure AI inference systems.

The challenge of enabling seamless, private, and secure access to data without disrupting the existing AI ecosystem and lifecycle is open and offers a fertile space for privacy technologies and the development of their use cases in fields such as healthcare, finance, and retail, to name a few. Such developments can also be applied to generic private computing applications and are not limited to AI or deep learning use cases.

We hope that the attention to the field continues fostering productive partnerships in the foreseeable future to address some of the challenges above and bring a future where the use of AI technology is compliant, safe, private, and explainable to fully realize the anticipated benefits of AI to societies and humanity worldwide.

ACKNOWLEDGMENTS

REFERENCES

[1] 2007–2022.   Partial   Dependence   and   Individual   Conditional   Expectation   plots. https://scikit-learn.org/stable/modules/partial_dependence.html.
[2] 2009. Netflix Prize. https://paperswithcode.com/dataset/netflix-prize.
[3] 2013. Health Insurance Portability and Accountability Act (HIPAA). https://www.hhs.gov/hipaa/index.html.
[4] 2016. General Data Protection Regulatory (GDPR). https://gdpr-info.eu.
[5] 2018. California Consumer Privacy Act (CCPA). http://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=
[6] 2019. The Next Generation of Data Sharing in Financial Services. https://www.weforum.org/whitepapers/the-next-generation-of-d
[7] 2020. https://www.inpher.io/news/2020/5/26/named-in-gartner-homomorphic-encryption-report.
[8] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *CCS'16*. 308–318.
[9] Nitin Agrawal, Reuben Binns, Max Van Kleek, Kim Laine, and Nigel Shadbolt. 2021. Exploring design and governance challenges in the development of privacy-preserving computation. In *CHI'21*.
[10] Avi Goldfarb Ajay Agrawal, Joshua Gans. 2018. Prediction Machines: The Simple Economics of Artificial Intelligence. https://www.predictionmachines.ai.
[11] Martin Albrecht, Melissa Chase, Hao Chen, Jintai Ding, Shafi Goldwasser, Sergey Gorbunov, Shai Halevi, Jeffrey Hoffstein, Kim Laine, Kristin Lauter, Satya Lokam, Daniele Micciancio, Dustin Moody, Travis Morrison, Amit Sahai, and Vinod Vaikuntanathan. 2018. *Homomorphic Encryption Security Standard*. Technical Report. https://homomorphicencryption.org/.
[12] Masaud Y. Alhassan, Daniel Günther, Ágnes Kiss, and Thomas Schneider. 2020. Efficient and Scalable Universal Circuits. *Journal of Cryptology (JoC)* (2020).

[13] Ahmad Al Badawi, Jack Bates, Flavio Bergamaschi, David Bruce Cousins, Saroja Erabelli, Nicholas Genise, Shai Halevi, Hamish Hunt, Andrey Kim, Yongwoo Lee, Zeyu Liu, Daniele Micciancio, Ian Quah, Yuriy Polyakov, Saraswathy R.V., Kurt Rohloff, Jonathan Saylor, Dmitriy Suponitsky, Matthew Triplett, Vinod Vaikuntanathan, and Vincent Zucca. 2022. OpenFHE: Open-Source Fully Homomorphic Encryption Library. Workshop on Encrypted Computing & Applied Homomorphic Cryptography. https://dualitytech.com/product/analytics-and-machine-learning/.
[14] Rajeev Balasubramonian, Jichuan Chang, Troy Manning, Jaime H Moreno, Richard Murphy, Ravi Nair, and Steven Swanson. 2014. Near-data processing: Insights from a MICRO-46 workshop. *IEEE Micro* 34, 4 (2014), 36–42.
[15] Mauro Barni, Pierluigi Failla, Vladimir Kolesnikov, Riccardo Lazzeretti, Ahmad-Reza Sadeghi, and Thomas Schneider. 2009. Secure Evaluation of Private Linear Branching Programs with Medical Applications. In *ESORICS'09*.
[16] Mauro Barni, Pierluigi Failla, Riccardo Lazzeretti, Ahmad-Reza Sadeghi, and Thomas Schneider. 2011. Privacy-Preserving ECG Classification with Branching Programs and Neural Networks. *TIFS'11*.
[17] Raef Bassily, Adam Smith, and Abhradeep Thakurta. 2014. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *FOCS'14*.
[18] Lejla Batina, Shivam Bhasin, Dirmanto Jap, and Stjepan Picek. 2019. CSI NN: Reverse Engineering of Neural Network Architectures Through Electromagnetic Side Channel. In *USENIX Security Symposium*. 515–532.
[19] Sebastian P. Bayerl, Tommaso Frassetto, Patrick Jauernig, Korbinian Riedhammer, Ahmad-Reza Sadeghi, Thomas Schneider, Emmanuel Stapf, and Christian Weinert. 2020. Offline Model Guard: Secure and Private ML on Mobile Devices. In *DATE'20*.
[20] Sarah Bird. 2022. Responsible AI investments and safeguards for facial recognition. https://azure.microsoft.com/en-us/blog/responsible-ai-investments-and-safeguards-for-facial-recognition/.
[21] Fabian Boemer, Rosario Cammarota, Daniel Demmler, Thomas Schneider, and Hossein Yalame. 2020. MP2ML: A Mixed-Protocol Machine Learning Framework for Private Inference. In *ARES'20*.
[22] Fabian Boemer, Anamaria Costache, Rosario Cammarota, and Casimir Wierzynski. 2019. nGraph-HE2: A High-Throughput Framework for Neural Network Inference on Encrypted Data. In *WAHC'19*.
[23] Fabian Boemer, Yixing Lao, Rosario Cammarota, and Casimir Wierzynski. 2019. nGraph-HE: A Graph Compiler for Deep Learning on Homomorphically Encrypted Data. In *CF'19*.
[24] Zvika Brakerski, Craig Gentry, and Vinod Vaikuntanathan. 2011. Fully Homomorphic Encryption without Boot-strapping. Cryptology ePrint Archive, Report 2011/277. https://eprint.iacr.org/2011/277.
[25] Luís TAN Brandão, Michael Davidson, and Apostol Vassilev. 2019. *Towards NIST Standards for Threshold Schemes for Cryptographic Primitives: A Preliminary Roadmap.* Technical Report. National Institute of Standards and Technology.
[26] Ferdinand Brasser, Tommaso Frassetto, Korbinian Riedhammer, Ahmad-Reza Sadeghi, Thomas Schneider, and Christian Weinert. 2018. VoiceGuard: Secure and Private Speech Processing. In *INTERSPEECH'18*.
[27] Ferdinand Brasser, David Gens, Patrick Jauernig, Ahmad-Reza Sadeghi, and Emmanuel Stapf. 2019. SANCTUARY: ARMing TrustZone with User-space Enclaves. In *NDSS'19*.
[28] Lennart Braun, Daniel Demmler, Thomas Schneider, and Oleksandr Tkachenko. 2022. MOTION – A Framework for Mixed-Protocol Multi-Party Computation. *Transactions on Privacy and Security* 25, 2 (2022).
[29] Thomas Brewster. 2020. Microsoft Urged To Follow Amazon And IBM: Stop Selling Facial Recognition To Cops After George Floyd's Death. https://www.forbes.com/sites/thomasbrewster/2020/06/11/microsoft-urged-to-follow-amazon-and-ibm-sto

[30] Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingy-ing Yang, Helen Toner, Ruth Fong, Tegan Maharaj, Pang Wei Koh, Sara Hooker, Jade Leung, Andrew Trask, Emma Bluemke, Jonathan Lebensold, Cullen O'Keefe, Mark Koren, Théo Ryffel, JB Rubinovitz, Tamay Besiroglu, Federica Carugati, Jack Clark, Peter Eckersley, Sarah de Haas, Maritza Johnson, Ben Laurie, Alex Ingerman, Igor Krawczuk, Amanda Askell, Rosario Cammarota, Andrew Lohn, David Krueger, Charlotte Stix, Peter Henderson, Logan Gra-ham, Carina Prunkl, Bianca Martin, Elizabeth Seger, Noa Zilberman, Seán Ó hÉigeartaigh, Frens Kroeger, Girish Sastry, Rebecca Kagan, Adrian Weller, Brian Tse, Elizabeth Barnes, Allan Dafoe, Paul Scharre, Ariel Herbert-Voss, Martijn Rasser, Shagun Sodhani, Carrick Flynn, Thomas Krendl Gilbert, Lisa Dyer, Saif Khan, Yoshua Bengio, and Markus Anderljung. 2020. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv:2004.07213 [cs.CY]
[31] Niklas Büscher, Daniel Demmler, Stefan Katzenbeisser, David Kretzmer, and Thomas Schneider. 2018. HyCC: Com-pilation of Hybrid Protocols for Practical Secure Computation. In *CCS'18*.
[32] Rosario Cammarota, Indranil Banerjee, and Ofer Rosenberg. 2018. Machine Learning IP Protection. In *ICCAD'18*.
[33] Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. 2020. Cryptanalytic Extraction of Neural Network Models. arXiv:2003.04884 [cs.LG]
[34] Nishanth Chandran, Divya Gupta, Aseem Rastogi, Rahul Sharma, and Shardul Tripathi. 2019. EzPC: Programmable and Efficient Secure Two-Party Computation for Machine Learning. In *EuroS&P'19*.

[35] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. 2011. Differentially private empirical risk minimization. *Journal of Machine Learning Research* 12, Mar (2011), 1069–1109.

[36] Lerong Chen, Jiawen Li, Yiran Chen, Qiuping Deng, Jiyuan Shen, Xiaoyao Liang, and Li Jiang. 2017. Accelerator-friendly neural-network training: Learning variations and defects in RRAM crossbar. In *DATE'17*.

[37] Eduardo Chielle, Oleg Mazonka, Homer Gamil, and Michail Maniatakos. 2022. Accelerating Fully Homomorphic Encryption by Bridging Modular and Bit-Level Arithmetic. In *ICCAD'22*. See also https://ia.cr/2018/1013.

[38] Ilaria Chillotti, Nicolas Gama, Mariya Georgieva, and Malika Izabachene. 2016. Faster fully homomorphic encryption: Bootstrapping in less than 0.1 seconds. In *ASIACRYPT'16*.

[39] Ilaria Chillotti, Marc Joye, and Pascal Paillier. 2021. Programmable Bootstrapping Enables Efficient Homomorphic Inference of Deep Neural Networks. In *Cyber Security Cryptography and Machine Learning*. Springer.

[40] Joseph Clements and Yingjie Lao. 2018. Backdoor attacks on neural network operations. In *GlobalSIP'18*. IEEE, 1154–1158.

[41] Joseph Clements and Yingjie Lao. 2018. Hardware trojan attacks on neural networks. *arXiv preprint arXiv:1806.05768* (2018).

[42] Janina Conboye. 2019. British Airways hit with record GBP 183M fine for data breach. https://www.ft.com/content/197a6758-a148-11e9-a282-2df48f366f7d.

[43] 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. arXiv preprint arXiv:1810.08130.

[44] Roshan Dathathri, Blagovesta Kostova, Olli Saarikivi, Wei Dai, Kim Laine, and Madan Musuvathi. 2020. EVA: An encrypted vector arithmetic language and compiler for efficient homomorphic computation. In *PLDI'20*.

[45] Roshan Dathathri, Olli Saarikivi, Hao Chen, Kim Laine, Kristin Lauter, Saeed Maleki, Madanlal Musuvathi, and Todd Mytkowicz. 2019. CHET: an optimizing compiler for fully-homomorphic neural-network inferencing. In *PLDI'19*.

[46] Kevin Deforth, Marc Desgroseilliers, Nicolas Gama, Mariya Georgieva, Dimitar Jetchev, and Marius Vuille. 2022. XORBoost: Tree Boosting in the Multiparty Computation Setting. *PETS* 2022, 4 (2022), 66–85. https://inpher.io/.

[47] Daniel Demmler, Thomas Schneider, and Michael Zohner. 2015. ABY - A Framework for Efficient Mixed-Protocol Secure Two-Party Computation. In *NDSS'15*.

[48] Ghada Dessouky, Farinaz Koushanfar, Ahmad-Reza Sadeghi, Thomas Schneider, Shaza Zeitouni, and Michael Zohner. 2017. Pushing the Communication Barrier in Secure Computation using Lookup Tables. In *NDSS'17*.

[49] Gaofeng Dong, Ping Wang, Ping Chen, Ruizhe Gu, and Honggang Hu. 2019. Floating-Point Multiplication Timing Attack on Deep Neural Network. In *SmartIoT'19*.

[50] Anuj Dubey, Rosario Cammarota, and Aydin Aysu. 2020. Bomanet: Boolean masking of an entire neural network. In *ICCAD'20*.

[51] Anuj Dubey, Rosario Cammarota, and Aydin Aysu. 2020. Maskednet: The first hardware inference engine aiming power side-channel protection. In *HOST'20*.

[52] Léo Ducas and Daniele Micciancio. 2014. FHEW: Bootstrapping Homomorphic Encryption in less than a second. Cryptology ePrint Archive, Report 2014/816. https://eprint.iacr.org/2014/816.

[53] Vasisht Duddu, Debasis Samanta, D Vijay Rao, and Valentina E. Balas. 2018. Stealing Neural Networks via Timing Side Channels. arXiv:1812.11720 [cs.CR]

[54] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*.

[55] European Commission. 2021. Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). https://digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificia

[56] Junfeng Fan and Frederik Vercauteren. 2012. Somewhat Practical Fully Homomorphic Encryption. Cryptology ePrint Archive, Report 2012/144. https://eprint.iacr.org/2012/144.

[57] Federal Trade Commission (FTC). 2019. FTC Imposes USD 5B Penalty and Sweeping New Privacy Restrictions on Facebook. https://www.ftc.gov/news-events/press-releases/2019/07/ftc-imposes-5-billion-penalty-sweeping-new-privacy-restric

[58] Food and Drug Administration (FDA). 2020. FDA Authorizes Marketing of First Cardiac Ultrasound Software That Uses Artificial Intelligence to Guide User. https://www.fda.gov/news-events/press-announcements/fda-authorizes-marketing-first-cardiac-ultrasound-software-uses-artifici

[59] Craig Gentry. 2009. *A Fully Homomorphic Encryption Scheme*. Stanford University PhD Thesis.

[60] Santosh Ghosh, Luis S Kida, Soham Jayesh Desai, and Reshma Lal. 2020. A >100 Gbps Inline AES-GCM Hardware Engine and Protected DMA Transfers between SGX Enclave and FPGA Accelerator Device. Cryptology ePrint Archive, Report 2020/178. https://eprint.iacr.org/2020/178.

[61] Ran Gilad-Bachrach, Nathan Dowlin, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy. In *ICML'16*.

[62] Oded Goldreich, Silvio Micali, and Avi Wigderson. 1987. How to Play any Mental Game. In *STOC'87*.

[63] Graphene Project. 2019. Graphene - A Library OS for Unmodified Applications. https://grapheneproject.io/.

[64] Daniel Günther, Ágnes Kiss, and Thomas Schneider. 2017. More Efficient Universal Circuit Constructions. In *ASI-ACRYPT'17*.

[65] Veneta Haralampieva, Daniel Rueckert, and Jonathan Passerat-Palmbach. 2020. A systematic comparison of encrypted machine learning solutions for image classification. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*. 55–59.

[66] Wilko Henecka, Stefan Kögl, Ahmad-Reza Sadeghi, Thomas Schneider, and Immo Wehrenberg. 2010. TASTY: Tool for Automating Secure Two-party Computations. In *CCS'10*.

[67] Sanghyun Hong, Pietro Frigo, Yiğitcan Kaya, Cristiano Giuffrida, and Tudor Dumitraş. 2019. Terminal brain damage: Exposing the graceless degradation in deep neural networks under hardware fault attacks. In *USENIX Security'19*. 497–514.

[68] Xing Hu, Ling Liang, Lei Deng, Shuangchen Li, Xinfeng Xie, Yu Ji, Yufei Ding, Chang Liu, Timothy Sherwood, and Yuan Xie. 2019. Neural Network Model Extraction Attacks in Edge Devices by Hearing Architectural Hints. arXiv:1903.03916 [cs.CR]

[69] Weizhe Hua, Zhiru Zhang, and G Edward Suh. 2018. Reverse Engineering Convolutional Neural Networks Through Side-channel Information Leaks. In *DAC'18*.

[70] Wenqin Huangfu, Lixue Xia, Ming Cheng, Xiling Yin, Tianqi Tang, Boxun Li, Krishnendu Chakrabarty, Yuan Xie, Yu Wang, and Huazhong Yang. 2017. Computation-oriented fault-tolerance schemes for RRAM computing systems. In *ASP-DAC'17*. IEEE, 794–799.

[71] Siam U Hussain and Farinaz Koushanfar. 2019. FASE: FPGA Acceleration of Secure Function Evaluation. In *FCCM'19*.

[72] Peter Hustinx. 2010. Privacy by Design: Delivering the Promises. *Identity in The Information Society* 3 (2010).

[73] Mohsen Imani, Saransh Gupta, Yeseong Kim, and Tajana Rosing. 2019. Floatpim: In-memory acceleration of deep neural network training with high precision. In *ISCA'19*. IEEE, 802–815.

[74] Mohsen Imani, Yeseong Kim, Sadegh Riazi, John Messerly, Patric Liu, Farinaz Koushanfar, and Tajana Rosing. 2019. A framework for collaborative learning in secure high-dimensional space. In *CLOUD'19*. IEEE, 435–446.

[75] Muhammad Ishaq, Ana L. Milanova, and Vassilis Zikas. 2019. Efficient MPC via Program Analysis: A Framework for Efficient Optimal Mixing. In *CCS'19*.

[76] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. 2020. High Accuracy and High Fidelity Extraction of Neural Networks. In *USENIX Security'20*. USENIX Association, Boston, MA.

[77] Chiraag Juvekar, Vinod Vaikuntanathan, and Anantha Chandrakasan. 2018. GAZELLE: A Low Latency Framework for Secure Neural Network Inference. In *USENIX Security'18*.

[78] Seny Kamara and Mariana Raykova. 2011. Secure Outsourced Computation in a Multi-Tenant Cloud. In *IBM Workshop on Cryptography and Security in Clouds*.

[79] Behnam Khaleghi, Mohsen Imani, and Tajana Rosing. 2020. Prive-HD: Privacy-Preserved Hyperdimensional Computing. arXiv:2005.06716 [cs.LG]

[80] Ágnes Kiss, Masoud Naderpour, Jian Liu, N Asokan, and Thomas Schneider. 2019. SoK: Modular and Efficient Private Decision Tree Evaluation. *PETS'19* (2019).

[81] Ágnes Kiss and Thomas Schneider. 2016. Valiant's Universal Circuit is Practical. In *EUROCRYPT'16*.

[82] Michael Klachko, Mohammad Reza Mahmoodi, and Dmitri Strukov. 2019. Improving noise tolerance of mixed-signal neural networks. In *IJCNN'19*. IEEE, 1–8.

[83] Vladimir Kolesnikov and Thomas Schneider. 2008. A Practical Universal Circuit Construction and Secure Evaluation of Private Functions. In *FC'08*.

[84] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. 2014. The CIFAR-10 Dataset. https://www.cs.toronto.edu/~kriz/cifar.html.

[85] Nishant Kumar, Mayank Rathee, Nishanth Chandran, Divya Gupta, Aseem Rastogi, and Rahul Sharma. 2020. CrypT-Flow: Secure TensorFlow Inference. In *S&P'20*.

[86] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications* 10, 1 (2019), 1–8.

[87] Yann LeCun and Corinna Cortes. 2010. MNIST Handwritten Digit Database. http://yann.lecun.com/exdb/mnist/.

[88] B. Liu, Hai Li, Yiran Chen, Xin Li, Qing Wu, and Tingwen Huang. 2015. Vortex: Variation-aware training for memristor X-bar. In *DAC'15*.

[89] Hanlin Liu, Yu Yu, Shuoyao Zhao, Jiang Zhang, Wenling Liu, and Zhenkai Hu. 2021. Pushing the Limits of Valiant's Universal Circuits: Simpler, Tighter and More Compact. In *CRYPTO'21*.

[90] Jian Liu, Mika Juuti, Yao Lu, and Nadarajah Asokan. 2017. Oblivious Neural Network Predictions via MiniONN Transformations. In *CCS'17*.

[91] Y. Liu, D. Dachman-Soled, and A. Srivastava. 2019. Mitigating Reverse Engineering Attacks on Deep Neural Networks. In *ISVLSI'19*. 657–662.

[92] Yannan Liu, Lingxiao Wei, Bo Luo, and Qiang Xu. 2017. Fault injection attack on deep neural network. In *ICCAD'17*. IEEE, 131–138.

[93] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[94] Dahlia Malkhi, Noam Nisan, Benny Pinkas, and Yaron Sella. 2004. Fairplay — A Secure Two-Party Computation System. (2004).

[95] Pratyush Mishra, Ryan Lehmkuhl, Akshayaram Srinivasan, Wenting Zheng, and Raluca Ada Popa. 2020. DELPHI: A Cryptographic Inference Service for Neural Networks. In *USENIX Security'20*.

[96] Payman Mohassel and Peter Rindal. 2018. ABY3: A Mixed Protocol Framework for Machine Learning. In *CCS'18*.

[97] Payman Mohassel and Yupeng Zhang. 2017. SecureML: A System for Scalable Privacy-preserving Machine Learning. In *S&P'17*.

[98] Clare Naden. 2019. It's all about trust. https://www.iso.org/news/ref2452.html.

[99] Hoda Naghibijouybari, Ajaya Neupane, Zhiyun Qian, and Nael Abu-Ghazaleh. 2018. Rendered Insecure: GPU Side Channel Attacks Are Practical. In *CCS'18*. 2139–2153.

[100] Hamid Nejatollahi, Saransh Gupta, Mohsen Imani, Tajana Simunic Rosing, Rosario Cammarota, and Nikil Dutt. 2020. CryptoPIM: In-memory Acceleration for Lattice-based Cryptographic Hardware. Cryptology ePrint Archive, Report 2020/276. https://eprint.iacr.org/2020/276.

[101] Antoinette Price. 2020. Achieving Trustworthy AI with Standards. https://etech.iec.ch/issue/2020-03/achieving-trustworthy-ai-with

[102] Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. 2019. Bit-flip attack: Crushing neural network with progressive bit search. In *Proceedings of the IEEE International Conference on Computer Vision*. 1211–1220.

[103] M. Sadegh Riazi, Kim Laine, Blake Pelton, and Wei Dai. 2019. HEAX: An Architecture for Computing on Encrypted Data. In *ASPLOS'19*.

[104] M Sadegh Riazi, Bita Darvish Rouani, and Farinaz Koushanfar. 2019. Deep Learning on Private Data. *S&P'19*.

[105] M Sadegh Riazi, Mohammad Samragh, Hao Chen, Kim Laine, Kristin Lauter, and Farinaz Koushanfar. 2019. XONN: XNOR-based Oblivious Deep Neural Network Inference. In *USENIX Security'19*.

[106] M Sadegh Riazi, Christian Weinert, Oleksandr Tkachenko, Ebrahim M Songhori, Thomas Schneider, and Farinaz Koushanfar. 2018. Chameleon: A Hybrid Secure Computation Framework for Machine Learning Applications. In *ASIACCS'18*.

[107] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *SIGKDD'16*.

[108] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-precision model-agnostic explanations. In *AAAI'18*.

[109] Mike Rosulek and Lawrence Roy. 2021. Three halves make a whole? Beating the half-gates lower bound for garbled circuits. In *CRYPTO'21*.

[110] Bita Darvish Rouhani, M Sadegh Riazi, and Farinaz Koushanfar. 2018. DeepSecure: Scalable Provably-Secure Deep Learning. In *DAC'18*.

[111] Benjamin IP Rubinstein, Peter L Bartlett, Ling Huang, and Nina Taft. 2009. Learning in a large function space: Privacy-preserving mechanisms for SVM learning. *arXiv preprint arXiv:0911.5708* (2009).

[112] Ahmad-Reza Sadeghi and Thomas Schneider. 2008. Generalized Universal Circuits for Secure Evaluation of Private Functions with Application to Data Classification. In *ICISC'08*.

[113] Ali Shafiee, Anirban Nag, Naveen Muralimanohar, Rajeev Balasubramonian, John Paul Strachan, Miao Hu, R Stanley Williams, and Vivek Srikumar. 2016. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars. *ACM SIGARCH Computer Architecture News* 44, 3 (2016), 14–26.

[114] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

[115] Karthee Sivalingam and Nina Mujkanovic. 2019. Graph Compilers for Artifical Intelligence Training and Inference. https://www.sodalite.eu/content/graph-compilers-ai-training-and-inference.

[116] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. 2013. Stochastic gradient descent with differentially private updates. In *2013 IEEE Global Conference on Signal and Information Processing*. IEEE, 245–248.

[117] E. M. Songhori, S. U. Hussain, A. Sadeghi, T. Schneider, and F. Koushanfar. 2015. TinyGarble: Highly Compressed and Scalable Sequential Garbled Circuits. In *S&P'15*.

[118] Aswin Sreedhar, Sandip Kundu, and Israel Koren. 2012. On reliability Trojan injection and detection. *Journal of Low Power Electronics* 8, 5 (2012), 674–683.

[119] Mark Tehranipoor, Waleed Khalil, Matthew Casto, Yousef Iskander, Brian Dupaix, Rosario Cammarota, and Brian Cohen. 2019. Trusted And Assured MicroElectronics (TAME) forum: Working Groups Report. https://dforte.ece.ufl.edu/wp-content/uploads/sites/65/2020/08/TAME-Report-FINAL.pdf .

[120] Shruti Tople, Karan Grover, Shweta Shinde, Ranjita Bhagwan, and Ramachandran Ramjee. 2018. Privado: Practical and Secure DNN Inference. *CoRR* abs/1810.00602 (2018). arXiv:1810.00602 http://arxiv.org/abs/1810.00602

[121] C. Torres-Huitzil and B. Girau. 2017. Fault and Error Tolerance in Neural Networks: A Review. *IEEE Access* 5 (2017).

[122] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction APIs. In *USENIX Security'16*.

[123] Florian Tramèr and Dan Boneh. 2019. Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware. In *ICLR'19*.

[124] Amos Treiber, Alejandro Molina, Christian Weinert, Thomas Schneider, and Kristian Kersting. 2020. CryptoSPN: Privacy-preserving Sum-Product Network Inference. In *ECAI'20*.

[125] Leslie G Valiant. 1976. Universal Circuits (Preliminary Report). In *STOC'76*.

[126] Bob Violino. 2022. Rising premiums, more restricted cyber insurance coverage poses big risk for companies. https://www.cnbc.com/2022/10/11/companies-are-finding-it-harder-to-get-cyber-insurance-.html .

[127] Wenhao Wang, Yichen Jiang, Qintao Shen, Weihao Huang, Hao Chen, Shuang Wang, XiaoFeng Wang, Haixu Tang, Kai Chen, Kristin Lauter, and Dongdai Lin. 2019. Toward Scalable Fully Homomorphic Encryption Through Light Trusted Computing Assistance. arXiv preprint arXiv:1905.07766.

[128] Lingxiao Wei, Bo Luo, Yu Li, Yannan Liu, and Qiang Xu. 2018. I Know What You See: Power Side-Channel Attack on Convolutional Neural Network Accelerators. In *ACSAC'18*.

[129] Zack Whittaker. 2019. After criticism, Homeland Security drops plans to expand airport face recognition scans to US citizens. https://techcrunch.com/2019/12/05/homeland-security-drops-airport-citizens-face-scans/ .

[130] Lixue Xia, Wenqin Huangfu, Tianqi Tang, Xiling Yin, Krishnendu Chakrabarty, Yuan Xie, Yu Wang, and Huazhong Yang. 2018. Stuck-at Fault Tolerance in RRAM Computing Systems. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 8, 1 (2018), 102–115.

[131] Lixue Xia, Mengyun Liu, Xuefei Ning, Krishnendu Chakrabarty, and Yu Wang. 2017. Fault-tolerant training with on-line fault detection for RRAM-based neural computing systems. In *DAC'17*.

[132] Xiaolin Xu, Fahim Rahman, Bicky Shakya, Apostol Vassilev, Domenic Forte, and Mark Tehranipoor. 2019. Electronics Supply Chain Integrity Enabled by Blockchain. *ACM Trans. Des. Autom. Electron. Syst.* (2019).

[133] Mengjia Yan, Christopher W. Fletcher, and Josep Torrellas. 2020. Cache Telepathy: Leveraging Shared Resource Attacks to Learn DNN Architectures. In *USENIX Security'20*.

[134] Andrew Chi-Chih Yao. 1986. How to Generate and Exchange Secrets. In *FOCS'86*.

[135] Shuoyao Zhao, Yu Yu, Jiang Zhang, and Hanlin Liu. 2019. Valiant's Universal Circuits Revisited: An Overall Improvement and a Lower Bound. In *ASIACRYPT'19*.