Privacy and security in AI

WHITE PAPER

Trustworthy artificial intelligence

10 July 2023

Lea Demelius*, Andreas Trügler*, Simone Kopeinik*, Sebastian Scher*, Tomislav Nad**, and Dominik Kowald*

* Know-Center GmbH

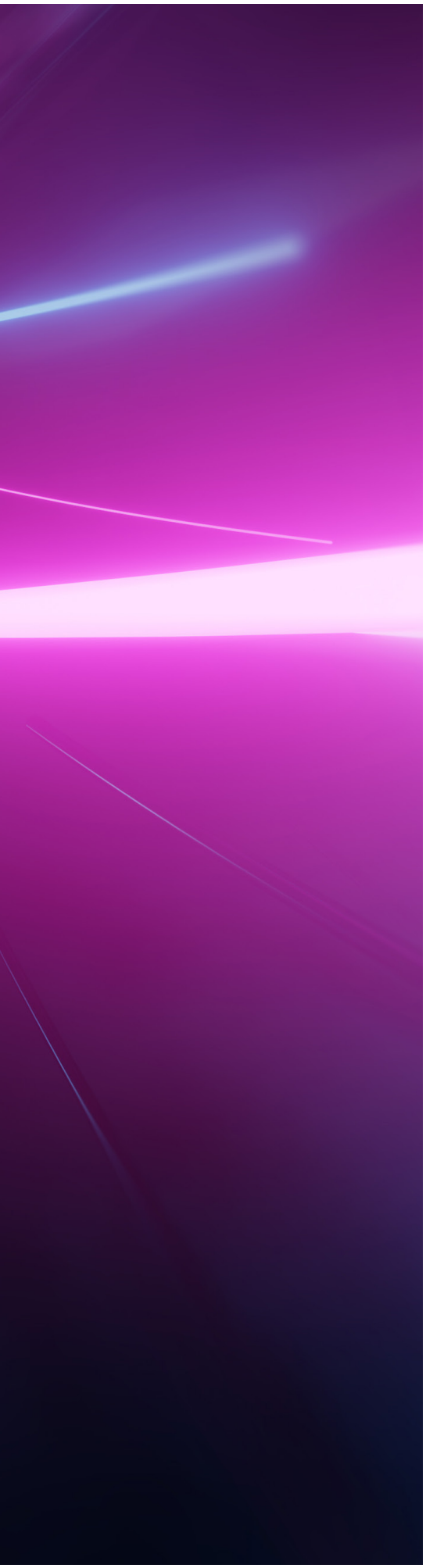
** SGS Digital Trusts Services GmbH

Partners of SGS

KNOW Center

SGS


Privacy and
security in AI



Contents:

1 Introduction

Security is a crucial aspect of the trustworthiness of an AI system and includes a broad range of topics. Privacy is a core component of information security and is linked to protecting personal and sensitive information. Both security and privacy concerns arise along the whole life cycle of AI systems, including data collection, model training, and model deployment and inference. Classical IT security aspects like secure data storage, secure transmissions or access control, have to be considered as well. However, in this white paper we focus on AI specific security and privacy aspects. For example, we take a look at malicious attacks on machine learning models (evasion, information extraction, poisoning, and backdoor attacks). Furthermore, in the past, data protection measures focused on protecting data during storage (i.e. data at rest) or transmission (data in transit). Thanks to recent advances in cryptography and privacy-enhancing technologies, new methods now also provide protection during computations (data in use). For AI models, this means that data and model can be protected against attacks during the training and inference process.

A lack of security privacy in AI systems can have serious consequences, ranging from data breaches and unauthorized access to sensitive information, as well as the manipulation of AI algorithms for malicious purposes and the perpetuation of biased decision-making. Additionally, compromised AI systems may lead to severe financial losses and reputational damage. They may even endanger human safety if AI is utilized in critical areas such as healthcare, autonomous vehicles or national security. Addressing these security and privacy concerns is crucial to foster trust in AI technology and ensure its responsible and ethical deployment across various domains.

Examples of past incidents of privacy and security violations can be found in the AI incident database[1]. For example, Facebook's friend suggestion feature recommended patients of a psychiatrist to each other (incident #406), OpenAI's generative model GPT-2 memorized and recited sensitive training data (incident #357) and suicide clips evaded TikTok's automated moderation system in a coordinated attack (incident #366).

Security and privacy have also become key aspects of new regulations, especially in the European Union. The legal and regulatory landscape in Europe regarding AI privacy and security has recently undergone significant developments. Key regulations include the General Data Protection Regulation (GDPR), which came into effect in 2018, the Data Governance Act and the proposed AI Act.

[1] *https://incidentdatabase.ai/*

AI model attacks may even endanger human safety if **AI** is utilized in critical areas such as healthcare, autonomous vehicles or national security.

AI model attacks in a nutshell

In order to secure an AI system, it is important to understand the threats it faces. Therefore, threat modeling, which defines the capabilities and knowledge of the attacker, as well as the attack goal and path, is an important first step. Typically, a distinction is made between attackers that can deviate from the agreed protocol (active/malicious) and those who try to learn as much as possible without violating the protocol (passive/semi-honest/honest-but-curious). Moreover, an attacker may be assumed to have finite or infinite computational power. Based on the attacker's knowledge, distinctions can be made between black-box attacks (which only access the model's output), white-box attacks (which access the full model), and gray-box attacks (which gain partial access). Attacks can be categorized into three types according to the attack goal (Bae et al. 2018, BSI 2023): (1) evasion attacks (2) poisoning and backdoor attacks (3) privacy attacks.

The goal of an evasion attack (including adversarial attacks) is to cause the AI model to misbehave or make incorrect predictions by providing input data specifically crafted to evade detection or classification. Defenses against evasion attacks range from traditional security measures, such as input validation and anomaly detection, to specialized techniques like adversarial training, defensive distillation and input preprocessing.

Poisoning attacks purposely cause malfunctioning or performance loss during model training. Backdoor attacks are targeted poisoning attacks that teach the model to produce a deliberate output change in response to a certain trigger. Defenses against poisoning and backdoor attacks include: using trusted sources for pre-trained models and training data, fine-tuning the model with trusted training samples, identifying triggers by testing if the model's output changes with small input changes, pruning the model, detecting the attack, using specifically designed regularization techniques and applying certified robustness (Jia et al. 2022).

Privacy attacks aim to reconstruct the model or (part of) its training data. The most common privacy attacks include:

- Membership inference attacks: with the aim of determining whether a sample was used for training
- Attribute inference attacks: with the aim of reconstructing sensitive attributes of individual records
- Model inversion attacks: with the aim of inferring features that characterize classes from the training data
- Model extraction/stealing attacks: with the aim of reconstructing the model's behavior, architecture and/or parameters

Defenses may include restricting the model output, sanitizing the training data, avoiding overfitting (e.g. by using regularization), and applying privacy-enhancing technologies (PETs).

Privacy-enhancing technologies (PETs)

There is a rising interest in PETs that can protect data in use. Each method is designed for a certain threat scenario and has its own advantages and limitations.

Fully homomorphic encryption (FHE) (Rechberger and Walch 2022, Smart 2015, Phong et al. 2018) allows computation on encrypted data. In the context of AI, FHE enables the use of cloud services (i.e. machine learning as a service) without disclosing the content of the data, or the access to pre-trained machine learning models, without disclosing the model properties. The main challenge of FHE is its computational overhead.

Secure multi-party computation (MPC) (Rechberger and Walch 2022, Evans et al. 2018) enables a group to jointly perform a computation without disclosing any participant's private inputs. Typical applications for machine learning include collaborative learning on combined datasets and private classification. MPC is communication intensive, especially with larger numbers of parties.

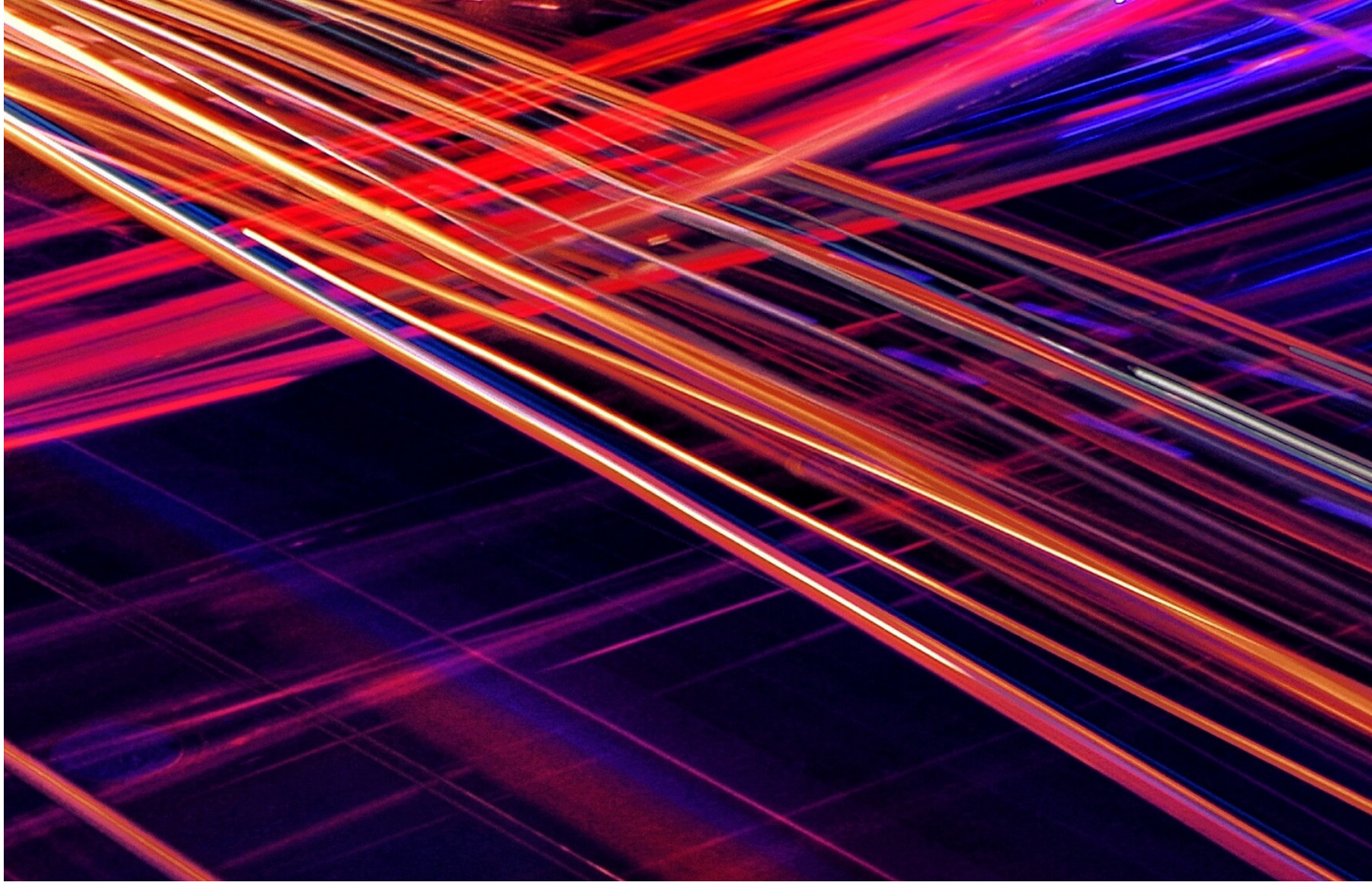
Differential privacy (DP) (Dwork 2008, Dwork and Roth 2014) is a mathematical notion used to quantify the risk of reconstructing the input data from the output of a computation. Traditionally, anonymization is applied to prevent such reconstruction, but there is an increasing awareness that despite the use of this strategy, re-identification is often still possible (Backstrom et al. 2007, Ganta et al. 2008, Narayanan and Shmatikov 2008). Differential privacy binds the maximum amount of information that the output of a computation discloses about an individual data point by adding curated noise to the computed function. The amount of added noise depends on the desired privacy level and the sensitivity of the function. In AI systems, noise can be added either to the input data, the objective function, the gradients or the output. For deep neural networks, gradient perturbation via differentially private stochastic gradient descent (DP-SGD) (Abadi et al. 2016) is the most common approach. DP has also been applied in the field of recommender systems, where a trade-off between a user's privacy and recommendation accuracy has been identified and addressed (Müllner et al. 2023). The main challenge of DP is the inherent trade-off between privacy and utility, as the noise inherently increases the uncertainty of the computation.

Federated learning (FL) (Zhang et al. 2021, Li et al 2020) is a machine learning method used to evaluate data from a large number of clients (e.g. mobile phones). Each client downloads the model, trains it locally, and shares the model updates with the central server which aggregates the updates from all clients. While this method provides some notion of security by keeping the private data local, training data can be reconstructed from the model updates (Yin et al. 2021).

Transfer learning is not a PET in the traditional sense but can indirectly contribute to privacy. In general, transfer learning refers to the fine-tuning of a pre-trained model on a new task for which only a small training dataset is available. For example, a publicly available pre-trained model can be fine-tuned for a specific task on a private training data set. This saves not only time and computational effort, but also decreases the amount of private data necessary to train a performant machine learning model. This allows, for example, local fine-tuning, where the private data never leaves the local storage.

Synthetic data approximates real data by retaining statistical properties, patterns and dependencies. Synthetic data generation is often seen as a promising solution for privacy-preserving data publishing/sharing, but the original data can often still be reconstructed (similar to anonymization techniques) (Stadler et al. 2022).

To provide increased privacy, the above methods can also be combined. For example, differential privacy can mitigate the risk of reconstruction in federated learning (Wei et al. 2020), transfer learning (Walch et al. 2022) and synthetic data (Tai et al. 2022, Stadler et al. 2022).



Metrics and methods to evaluate privacy and security

The vulnerability of AI models to privacy and security attacks can be evaluated through a combination of two complementary approaches: mathematical analysis and attack-based evaluation. Mathematical analysis provides formal guarantees, while attack-based evaluation offers real-world insights into how the model behaves under different attack scenarios.

Mathematical analysis can be used to prove and/or quantify specific security and privacy properties of a system. This approach is often applied in cryptography and can provide formal guarantees about the security of a system under certain assumptions. To make sure that the guarantee holds true, the implementation should be checked for errors and suitable parameter selection. Mathematical analysis is of particular relevance when a new security/privacy method is published.

Attack-based evaluation empirically measures if (and how much) a specific AI model is vulnerable to a specific attack. It assesses how susceptible the model is to different attack strategies and quantifies the model's robustness. Various metrics are commonly used to evaluate the performance of the model under attack, such as the success rate of the attack, the number of iterations needed for a successful attack, the attack accuracy and the minimal data changes needed for a successful attack (BSI 2022). Which metric(s) should be applied in which case depends on the type of attack used for evaluation and the assumptions made about the attacker's capabilities and knowledge. This, in turn, needs to be identified for the specific use case and model at hand through analysis of possible threat scenarios and by searching related literature. It is essential to recognize the limitations of attack-based evaluation. While it can identify vulnerabilities and weaknesses in the model, it does not provide an overall privacy/security guarantee. Moreover, it only covers specific attack scenarios tested during evaluation and may not account for unforeseen or sophisticated attacks.

Attack-based evaluation empirically measures if (and how much) a specific **AI** model is vulnerable to a specific attack.

Open issues and challenges

The field of AI security and privacy faces several open issues and challenges that require research and innovation. Despite the development of defense mechanisms, AI models still face challenges in achieving robustness against adversarial attacks. Additionally, the limitations and possible negative impacts of defense methods (e.g. decreased utility or computational overhead) must be taken into account. Researchers continue to explore more effective defenses to mitigate the impact of adversarial examples. The transferability of attacks across models poses a significant threat, necessitating more robust defense strategies. Defending against poisoning attacks and ensuring the integrity of training data is a critical challenge. Detecting and mitigating poisoned data inputs requires innovative techniques and strategies.

Preserving privacy in deep learning models remains a complex challenge, particularly when dealing with sensitive data. Methods like differential privacy show promise but need further investigation to strike the right balance between privacy protection and utility. Seamlessly integrating privacy-enhancing technologies (PETs) with existing AI systems requires research into compatibility, performance overheads, and ensuring that privacy is not compromised in the process. Many privacy and security-enhancing techniques can be computationally expensive, making them challenging to implement in resource-constrained environments like edge devices. Enabling secure model sharing and collaboration between multiple parties while preserving privacy remains a challenging problem.

Developing standardized metrics for evaluating privacy and security across different AI systems is essential to facilitate comparative analysis and benchmarking.

The emergence of advanced AI-based attacks, including the potential for AI systems to be used as attackers, raises new concerns. Preparing for the future landscape of adversarial AI is a critical challenge.

Addressing these open issues and challenges requires a collective effort from researchers, developers, policymakers, and organizations. Continuous research, collaboration and sharing of best practices are essential to create a more secure and privacy-respecting AI ecosystem.

Summary and outlook

Security and privacy are crucial aspects of AI systems and involve a broad range of different topics, ranging from traditional IT security aspects like secure data storage and access control, to AI specific attacks like evasion and privacy attacks. In this white paper we are focusing on the latter and provide an introduction to those threats and defenses. Evaluating the security of AI systems and models is challenging. Mathematical analysis provides formal guarantees and quantification of security and privacy properties, while attack-based evaluation offers real-world insights into model behavior under various attack scenarios. These evaluations are crucial in identifying weaknesses and implementing effective defense mechanisms to protect against potential threats.

New technologies have emerged in addition to standard software and systems security measures, including homomorphic encryption, secure multi-party computation, differential privacy and federated learning. Those technologies solve some of the issues and have become more industry-ready in recent years, e.g. due to improvement of the performance trade-offs, but they still have limitations and require further research and innovation.

Privacy and security in AI is an active research field and a central topic in the ongoing pursuit of trustworthy AI. Legal frameworks are in the process of being established and will influence further developments in the field. As AI continues to play a pivotal role in various domains, addressing these security and privacy concerns remains paramount to ensure responsible and ethical deployment, and to foster trust in this transformative technology.

Acknowledgements

Know-Center is a leading European research center for Big Data, Artificial Intelligence (AI) and Data-Driven business models. Know-Center is a COMET Centre within COMET – Competence Centers for Excellent Technologies. This program is funded by the Austrian Federal Ministries for Climate Policy, Environment, Energy, Mobility, Innovation and Technology (BMK), and for Labor and Economy (BMAW), represented by Österreichische Forschungsförderungsgesellschaft mbH (FFG), Steirische Wirtschaftsförderungsgesellschaft mbH (SFG) and the Province of Styria, Wirtschaftsagentur Vienna and Standortagentur Tyrol GmbH.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016, October). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security (pp. 308-318).

Backstrom, L., Dwork, C., & Kleinberg, J. (2007, May). Wherefore art thou R3579X? Anonymized social networks, hidden patterns, and structural steganography. In Proceedings of the 16th international conference on World Wide Web (pp. 181-190).

Bae, H., Jang, J., Jung, D., Jang, H., Ha, H., Lee, H., & Yoon, S. (2018). Security and privacy issues in deep learning. arXiv preprint arXiv:1807.11655.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends® in Theoretical Computer Science, 9(3–4), 211-407.

Dwork, C. (2008). Differential privacy: A survey of results. In Theory and Applications of Models of Computation: 5th International Conference, TAMC 2008, Xi'an, China, April 25-29, 2008. Proceedings 5 (pp. 1-19). Springer Berlin Heidelberg.

Evans, D., Kolesnikov, V., & Rosulek, M. (2018). A pragmatic introduction to secure multi-party computation. Foundations and Trends® in Privacy and Security, 2(2-3), 70-246.

Federal Office for Information Security (BSI), 2022. Security of AI-Systems: Fundamentals – Adversarial Deep Learning. *https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Security-of-AI-systems_fundamentals.pdf?__blob=publicationFile&v=4*

Federal Office for Information Security (BSI), 2023. AI security concerns in a nutshell – Practical AI-Security guide. *https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/KI/Practical_AI-Security_Guide_2023.html*

Ganta, S. R., Kasiviswanathan, S. P., & Smith, A. (2008, August). Composition attacks and auxiliary information in data privacy. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 265-273).

Jia, J., Liu, Y., Cao, X., & Gong, N. Z. (2022). Certified Robustness of Nearest Neighbors against Data Poisoning and Backdoor Attacks. Proceedings of the AAAI Conference on Artificial Intelligence, 36(9), 9575-9583. *https://doi.org/10.1609/aaai.v36i9.21191*

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE signal processing magazine, 37(3), 50-60.

Müllner, P., Schedl, M., Lex, E., & Kowald, D. (2023). ReuseKNN: Neighborhood Reuse for Privacy-Aware Recommendations. ACM Transactions on Intelligent Systems and Technology.

Narayanan, A., & Shmatikov, V. (2008, May). Robust de-anonymization of large sparse datasets. In 2008 IEEE Symposium on Security and Privacy (sp 2008) (pp. 111-125). IEEE.

Phong, L. T., Aono, Y., Hayashi, T., Wang, L., & Moriai, S. (2018). Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. IEEE Trans. Information Forensics and Security, 13(5), pp. 1333-1345.

Rechberger, C., & Walch, R. (2022). Privacy-preserving machine learning using cryptography. In Security and Artificial Intelligence: A Crossdisciplinary Approach (pp. 109-129). Cham: Springer International Publishing.

Smart, P. N. (2015). Cryptography made simple. Springer International Publishing Switzerland.

Stadler, T., Oprisanu, B., & Troncoso, C. (2022). Synthetic data – anonymisation groundhog day. In 31st USENIX Security Symposium (USENIX Security 22) (pp. 1451-1468).

Tai, B. C., Li, S. C., Huang, Y., & Wang, P. C. (2022, November). Examining the Utility of Differentially Private Synthetic Data Generated using Variational Autoencoder with TensorFlow Privacy. In 2022 IEEE 27th Pacific Rim International Symposium on Dependable Computing (PRDC) (pp. 236-241). IEEE.

Walch, R., Sousa, S., Helminger, L., Lindstaedt, S., Rechberger, C., & Trügler, A. (2022). CryptoTL: Private, efficient and secure transfer learning. arXiv preprint arXiv:2205.11935.

Wei, K., Li, J., Ding, M., Ma, C., Yang, H. H., Farokhi, F., ... & Poor, H. V. (2020). Federated learning with differential privacy: Algorithms and performance analysis. IEEE Transactions on Information Forensics and Security, 15, 3454-3469.

Yin, H., Mallya, A., Vahdat, A., Alvarez, J. M., Kautz, J., & Molchanov, P. (2021). See through gradients: Image batch recovery via gradinversion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 16337-16346).

Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. Knowledge-Based Systems, 216, 106775.

Privacy and Security in **AI**

SGS.COM/DIGITAL

CONTACT US

SGS

Emerging Technology

✉ Enquiry.Emerging-Technology@sgs.com

KNOW CENTER

Leading Research and Innovation Center for Trustworthy AI

https://know-center.at/

✉ info@know-center.at

WHEN YOU NEED TO BE SURE