Self-Portrayals of GI Junior Fellows

Franziska Boenisch*

A self-portrayal of GI Junior Fellow Franziska Boenisch: trustworthy machine learning for individuals

https://doi.org/10.1515/itit-2024-0101
Received December 15, 2024; accepted February 19, 2025;
published online April 2, 2025

Abstract: Machine learning (ML) is increasingly deployed in critical domains such as healthcare, finance, and autonomous driving, where the use of sensitive data raises significant privacy challenges. My research places individuals and their data at the center of ML privacy, building systems that protect individuals' privacy without sacrificing performance. I focus on (1) exploring the threat space in ML privacy to inspire targeted protection, (2) analyzing the root cause of privacy leakage from ML models, and (3) developing individualized privacy guarantees that protect data according to individuals' unique needs while improving privacy-utility trade-offs. My vision is to advance privacy-preserving ML to address the evolving challenges of increasingly complex ML models and systems. As models grow in scale, integrate diverse data modalities, and become embedded in critical societal applications, protecting individual privacy becomes both more urgent but also more challenging. My goal is to create methods that ensure privacy across a broad spectrum of ML applications, while also addressing the interplay between privacy and other trustworthy ML aspects, and aligning technical privacy measures with legal and societal expectations to meet individual rights.

Keywords: trustworthy machine learning; privacy; GI junior fellow

1 Introduction

Machine learning (ML) is increasingly deployed in critical domains such as healthcare, finance, and autonomous driving, where the use of sensitive data raises significant privacy challenges. Regulatory frameworks, including the European Union's General Data Protection Regulation (GDPR), the United States' California Consumer Privacy Act (CCPA), and the Health Insurance Portability and Accountability Act (HIPAA), have been established to address privacy concerns. However, their application to ML systems remains limited due to fundamental gaps in understanding the ML-specific threat space, insufficient insights into the root causes of privacy leakage, and inadequate consideration of individuals' heterogeneous privacy requirements when implementing protective measures. These unresolved issues do not only increase the risk of privacy violations but also threaten to undermine the adoption of ML in socially critical domains.

My research addresses these gaps by **developing privacy frameworks that prioritize the protection of individuals and their data in ML systems**. To achieve this, I focus on three key directions. First, I **explore novel threat spaces in ML**, identifying how individuals' data may be exposed through previously unconsidered vulnerabilities. By broadening the understanding of potential risks, my work inspires protection against unintended leakage. Second, I **analyze the root causes of privacy risks in ML**, with a focus on how memorization and other model behaviors lead to data leakage. By uncovering the underlying mechanisms behind these risks, my research offers valuable insights for designing targeted mitigation strategies. This work emphasizes the importance of implementing privacy-by-design principles, enabling a more effective balance between robust privacy protection and optimal model performance. Finally, I **develop protection measures with individualized privacy guarantees**. Unlike traditional approaches that rely on protecting all training data points of an ML model uniformly, these individualized guarantees allow to protect every data point according to its unique privacy requirements or preferences. This approach does not only put individuals into focus, but by leveraging the training data more effectively, also yields better privacy-utility trade-offs.

My work advances privacy research in ML by **prioritizing an individual-centric approach**, enabling

***Corresponding author: Franziska Boenisch**, CISPA, Saarbrücken, Germany, E-mail: boenisch@cispa.de.
https://orcid.org/0000-0002-2111-2234

systems that respect individualized privacy without sacrificing performance.

2 Identifying novel ML privacy risks

To date, the awareness of potential privacy risks in ML is low, even among ML practitioners, *i.e.*, individuals who are in charge of designing and implementing ML systems [1]. Additionally, methods to protect ML privacy are significantly less widely known than methods protecting other aspects, such as ML security. Finally, large parts of the threat space against ML privacy still remain to be explored.

In my work [2], I discovered the inherent privacy risks of Federated Learning (FL), currently the most widely applied framework for decentralized ML. My findings invalidate a large body of literature claiming that FL provides a non-formal notion of privacy resulting from the fact that instead of raw user data, only ML model gradients of this data are exchanged. These prior beliefs were mainly based on the assumptions that reconstruction methods are computationally costly, and typically obtain low-fidelity reconstructions, especially for data that is high dimensional or contains multiple instances from the same class, or when the local gradients are calculated over many data points inside a mini-batch. My work disproves these claims and highlights that even **individual users' gradients calculated on large data mini-batches and high-dimensional complex datasets directly leak individual training data points**. Since, in FL, these user gradients are directly sent to a central party coordinating the distributed learning process, this allows for direct and perfect, *i.e.*, **zero-error data extraction at near-zero computation costs**, resulting in severe privacy breaches. I also showed that passive privacy leakage can be significantly increased through a novel *active* data extraction attack which inconspicuously manipulates the weights of the shared model sent out to the users. This demonstrates that FL, often used in highly sensitive contexts across the industry, such as in Google's mobile keyboard prediction or Apple's news personalization, is not private and relies on an implicit trust assumption on the central party.

In follow-up work [3], I showed that the privacy risks still persist even when FL is hardened by distributed Differential Privacy (DP) and secure aggregation – previously considered the most private yet practical version of the protocol. By instantiating a simple sybil attack where the server introduces a few manipulated devices into training, I demonstrated that individual users do not necessarily obtain their promised privacy guarantees. This is because distributed DP relies on every user adding a share of the noise to implement protection against leakage. Yet, users do not have any guarantees that other users (potentially sybil devices) would add the right amounts of noise, leaving their data unprotected. Having each user add the necessary amount of noise locally does not offer a practical alternative as it deteriorates the final model's performance significantly. Thereby, my work showed that prior solutions to improve FL privacy *either* provide utility *or* provable privacy protection, but not both. My insights also revealed that the power-imbalance between the server and the users is the main cause for FL's vulnerability: the server controls the shared model, can introduce devices into the protocol, and can sample participants for given training rounds while the users have no guarantees on the training protocol, the other users and their data. This inspired the design of novel solutions towards truly private FL. To date, my initial paper [2] has been cited more than 190 times and inspired multiple follow up works that achieve actual privacy protection while maintaining model performance by shifting power from the server to the users, e.g. [4]. The line of work also sparked wide interest in the industry. For example, I was invited to present my findings at Microsoft Research, Apple, and Google.

Following my presentation, Google revised its official stance on FL, shifting from describing it as *privacy-preserving* to framing it as *data minimizing*.

3 Understanding the causes of privacy leakage in ML

In my work, I also analyzed the root causes of privacy leakage in ML through the framework of *memorization*. Memorization, defined as the retention of specific training data by ML models, is often linked to privacy risks, especially in generative models where it can lead to the unintended full exposure of sensitive data during inference (for example, when a diffusion model generates a 1:1 copy of its training data points). While prior research on memorization has primarily focused on supervised learning, self-supervised learning – central to modern large language and vision models – has received far less attention and was considered only empirically. Traditional formal definitions of memorization, centered on predicting labels, fail to address these models' high-dimensional outputs. To bridge this gap, my research has formalized and examined memorization in self-supervised vision and generative models like diffusion models.

In [5], together with my first PhD student, I introduced the first formal definition of memorization tailored for self-supervised vision encoders. Our definition operates solely on the encoders' output representations, making it universally applicable across state-of-the-art self-supervised pretraining frameworks. It captures memorization by identifying a significantly higher representation alignment on memorized data – where encoders produce notably more similar representations for different augmented views of memorized input images. This novel lens allowed us to analyze memorization across diverse encoders and pretraining methods, uncovering that, akin to supervised learning, self-supervised learning inherently requires a degree of memorization to generalize effectively to downstream tasks. Surprisingly, we found that this holds even when the pretraining data comes from a different distribution than the downstream data: memorizing data from one distribution enhances the encoder's ability to perform tasks on others. Our theoretical results suggest that memorizing outlier and atypical data creates a more structured latent space, which supports this generalization. Beyond traditional classification tasks, we demonstrated that this phenomenon extends to more complex tasks such as segmentation and depth estimation.

Overall, our insights underscore the inevitability of trade-offs between privacy and generalization, emphasizing the need to design methods that achieve desirable trade-offs.

In our follow-up work [6], we developed a novel method to *localize* memorization within the parameters of self-supervised vision encoders. Specifically, we introduced measures to identify memorization at the level of both individual layers and individual neurons, enabling us to **pinpoint specific neurons that memorize individual training data points in various encoders**. Our localization techniques allowed us to systematically compare memorization behaviors across different self-supervised frameworks and encoder architectures and their supervised learning counterparts. While supervised learning predominantly assigns responsibility for *classes* to individual neurons, we found that self-supervised learning distributes memorization of *individual data points* in a large number of neurons. We attribute the contrast to the differing training objectives: supervised learning focuses on class distinctions, whereas self-supervised learning distinguishes between individual instances. Our findings carry significant privacy implications, revealing that self-supervised models initially expose more information about individual data points. However, we also demonstrated that this risk can be mitigated by pruning neurons responsible for memorization, offering a practical pathway to enhancing privacy in self-supervised models.

Building on these insights, we extended our research to the more complex setting of generative diffusion models [7], where memorization can result in the direct reproduction of training data. To address this, we developed a computationally efficient localization scheme tailored to the large architectures and iterative generation processes characteristic of these models. Remarkably, we discovered that even in such complex systems, individual neurons can be responsible for memorizing specific data points – and in some cases, a single neuron can encode multiple data points. By selectively dampening or deactivating these neurons, we demonstrated that memorization could be effectively mitigated without compromising model performance. Our approach offers a practical and scalable solution to privacy risks in generative models. Crucially, it is minimally invasive and applicable to already trained models, eliminating the need for costly retraining. This ensures that both the computational investment in training and the utility of the model are preserved, making it a highly attractive option for addressing privacy concerns in generative models.

These studies highlight the dual role of memorization in driving model performance and contributing to privacy risks. Our work provides key insights into the memorization behavior of self-supervised and generative models and the associated privacy leakage. By localizing memorization, we enable targeted mitigation strategies that address privacy concerns while maintaining model utility.

4 Providing individualized privacy guarantees

In traditional privacy-preserving ML, a single privacy level, often expressed as a privacy parameter ε, governs the privacy guarantees for the entire dataset. In the mathematical framework of differential privacy (DP), this parameter upper-bounds the potential privacy leakage, with higher values of ε indicating greater leakage and lower values indicating stronger privacy protection. However, this uniform approach fails to account for the *diverse privacy preferences or requirements* of individuals in the dataset. As a result, in ML, to avoid violating anyone's privacy preferences, the strictest privacy requirement encountered in the training dataset dictates the overall privacy level. As stricter privacy requirements rely on the addition of larger amounts of noise during training, this degrades model performance.

To enhance model performance while respecting individuals' privacy preferences, I pioneered the concept of individualized privacy in machine learning (ML). This groundbreaking approach tailors privacy guarantees to the specific needs of each individual, overcoming the limitations of traditional one-size-fits-all methods. I developed both a theoretical framework and practical algorithms to train ML models with individualized per-data point privacy guarantees. In [8], I extended the PATE algorithm [9], one of the two canonical algorithms for private ML, to support individualized privacy. Specifically, I modified the noisy knowledge transfer mechanism at the algorithm's core to allow data points to contribute varying amounts of information, resulting in tailored privacy guarantees. Similarly, in [10], I introduced two novel individualized extensions to the widely-used DPSGD algorithm [11]. These extensions adjust either the likelihood of individual data points being included in training or the amount of information transmitted through gradients, depending on their privacy preferences. This ensures that each data point's unique privacy requirements are met while introducing no computational overhead in comparison to non-individualized training. Both works include rigorous theoretical proofs verifying the privacy guarantees and experimental results showing that individualized privacy can improve the privacy-utility trade-offs: by tailoring privacy levels, these methods utilize data more effectively, enabling better model performance while respecting individual privacy needs.

This work offers a new approach to privacy-preserving machine learning by tailoring privacy guarantees to individual needs. It addresses the limitations of traditional privacy frameworks and improves the balance between privacy and utility. As a result, these methods enhance the applicability and effectiveness of privacy-preserving ML in a variety of real-world scenarios.

5 Ongoing and future work

My research vision focuses on *integrating privacy into trustworthy ML systems with a strong emphasis on individual protection.* To achieve this, I have identified three key goals and specific project directions that will drive my future work and vision.

5.1 Goal 1: privacy in the age of foundation models

The rise of foundation models, which undergo a two-step process of *pretraining* and *fine-tuning*, presents new privacy challenges, particularly due to the complex interaction between pretraining and fine-tuning data. To address these challenges, I am focusing on the following key directions: **Direction 1: Privacy in Foundation Models.** My goal is to investigate the novel privacy risks inherent in the pretraining and fine-tuning process. By thoroughly exploring this emerging threat landscape, identifying potential sources of leakage, and developing targeted, effective mitigation methods, I aim to ensure that foundation models can be applied safely in sensitive domains, thereby benefiting society. **Direction 2: Privacy Across Modalities.** In addition, I am interested in exploring privacy risks that extend beyond single modalities. Many modern foundation models process multiple modalities, such as combining text, images, or even video. As a result, information about individuals may be distributed across these modalities. I seek to understand how privacy can be assessed and protected across modalities, and to develop effective defenses that preserve individuals' privacy in these complex settings.

5.2 Goal 2: privacy and its interplay with other aspects of trustworthy ML

My second major goal is to explore the tensions and synergies between privacy and other aspects of trustworthy ML, such as fairness and robustness, and to leverage these synergies for better-performing systems. **Direction 1, Analyzing Trade-Offs between Privacy and Other Trustworthy Aspects:** I will continue [12], [13] to characterize the trade-offs between privacy and other aspects of trustworthy ML, such as fairness, robustness, and model utility. By understanding these trade-offs more deeply, I aim to identify how to mitigate negative impacts while maintaining privacy guarantees. **Direction 2, Joint Optimization:** In preliminary work [14], [15], we are exploring methods for jointly optimizing privacy, fairness, and utility, using approaches such as game theory to find Pareto frontier solutions that balance these aspects. Extending this line of work, my goal is to design systems where privacy of individuals does not have to come at the cost of fairness or performance.

5.3 Goal 3: governance and legal alignment for privacy

Effective privacy governance ensures model deployment aligns with privacy standards and regulations by establishing clear documentation, monitoring, and accountability throughout the model lifecycle. **Direction 1, Privacy Governance and Auditing Frameworks:** I aim to develop governance frameworks that provide comprehensive oversight of privacy practices in ML systems. These frameworks will identify and mitigate privacy risks, particularly concerning sensitive data exposure and model memorization, while offering tools to monitor real-world deployment impacts and maintain compliance. **Direction 2, Alignment with**

Legal and Regulatory Standards: Building on my experiences at the intersection of technical and legal privacy risk assessment [16], I plan to design tools ensuring ML systems comply with regulations like the GDPR and support privacy audits for legal contexts, enabling organizations to navigate complex regulatory landscapes.

Overall, my efforts will lead to more socially reliable ML systems that not only achieve high performance but also respect and protect the privacy of individuals at every step of their lifecycle.

Research ethics: Not applicable.
Informed consent: Not applicable.
Author contributions: The author has accepted responsibility for the entire content of this manuscript and approved its submission.
Use of Large Language Models, AI and Machine Learning Tools: None declared.
Conflict of interest: The author states no conflict of interest.
Research funding: None declared.
Data availability: Not applicable.

References

[1] F. Boenisch, V. Battis, N. Buchmann, and M. Poikela, ""I never thought about securing my machine learning systems": a study of security and privacy awareness of machine learning practitioners," in *Mensch und Computer 2021*, 2021, pp. 520–546.

[2] F. Boenisch, A. Dziedzic, R. Schuster, A. S. Shamsabadi, I. Shumailov, and N. Papernot, "When the curious abandon honesty: federated learning is not private," in *8th IEEE European Symposium on Security and Privacy (EuroS&P '23)*, 2023.

[3] F. Boenisch, A. Dziedzic, R. Schuster, A. S. Shamsabadi, I. Shumailov, and N. Papernot, "Reconstructing individual data points in federated learning hardened with differential privacy and secure aggregation," in *8th IEEE European Symposium on Security and Privacy (EuroS&P '23)*, 2023.

[4] N. Franzese, *et al*., "Robust and actively secure serverless collaborative learning," *Adv. Neural Inf. Process. Syst.*, vol. 36, 2024.

[5] W. Wang, M. Ahmad Kaleem, A. Dziedzic, M. Backes, N. Papernot, and F. Boenisch, "Memorization in self-supervised learning improves downstream generalization," in *The Twelfth International Conference on Learning Representations (ICLR)*, 2023.

[6] W. Wang, A. Dziedzic, M. Backes, and F. Boenisch, "Localizing memorization in ssl vision encoders," in *Accepted for: Advances in Neural Information Processing Systems (NeurIPS)*, vol. 38, 2024.

[7] D. Hintersdorf, L. Struppek, K. Kersting, A. Dziedzic, and F. Boenisch, "Finding nemo: localizing neurons responsible for memorization in diffusion models," in *Accepted for: Advances in Neural Information Processing Systems (NeurIPS)*, vol. 38, 2024.

[8] F. Boenisch, C. Mühl, R. Rinberg, J. Ihrig, and A. Dziedzic, "Individualized pate: differentially private machine learning with individual privacy guarantees," in *23rd Privacy Enhancing Technologies Symposium (PoPETs)*, 2023.

[9] N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," in *International Conference on Learning Representations*, 2016.

[10] F. Boenisch, C. Mühl, A. Dziedzic, R. Rinberg, and N. Papernot, "Have it your way: individualized privacy assignment for dp-sgd," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 37, 2023.

[11] M. Abadi, *et al*., "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016, pp. 308–318.

[12] A. Kowalczuk, *et al*., "Benchmarking robust self-supervised learning across diverse downstream tasks," in *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

[13] J. Wu, A. A. Ghomi, D. Glukhov, J. C. Cresswell, F. Boenisch, and N. Papernot, "Augment then smooth: reconciling differential privacy with certified robustness," *Trans. Mach. Learn. (TMLR)*, 2024. https://openreview.net/forum?id=YN0IcnXqsr.

[14] M. Yaghini, P. Liu, F. Boenisch, and N. Papernot, "Learning to walk impartially on the pareto frontier of fairness, privacy, and utility," in *NeurIPS 2023 Workshop on Regulatable ML*, 2023.

[15] M. Yaghini, P. Liu, F. Boenisch, and N. Papernot, "Regulation games for trustworthy machine learning," in *NeurIPS 2023 Workshop on Regulatable ML*, 2023.

[16] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, "A unified framework for quantifying privacy risk in synthetic data," in *23rd Privacy Enhancing Technologies Symposium (PoPETs)*, 2023.

Bionote

Franziska Boenisch
CISPA, Saarbrücken, Germany
boenisch@cispa.de
https://orcid.org/0000-0002-2111-2234

Franziska is a tenure-track faculty at the CISPA Helmholtz Center for Information Security where she co-leads the SprintML lab. Before, she was a Postdoctoral Fellow at the University of Toronto and Vector Institute advised by Prof. Nicolas Papernot. Her current research centers around private and trustworthy machine learning. Franziska obtained her Ph.D. at the Computer Science Department at Freie University Berlin, where she pioneered the notion of individualized privacy in machine learning. During her Ph.D., Franziska was a research associate at the Fraunhofer Institute for Applied and Integrated Security (AISEC), Germany. She received a Fraunhofer TALENTA grant for outstanding female early career researchers, the German Industrial Research Foundation prize for her research on machine learning privacy, and the Fraunhofer ICT Dissertation award 2023, and was named a GI-Junior Fellow in 2024.