

Nils Lukas

Assistant Professor • MBZUAI • Abu Dhabi, UAE
nils.lukas@mbzuai.ac.ae • nilslukas.github.io • [Gscholar](https://scholar.google.com/citations?user=HgkQWgIAAAJ&hl=en)
Updated on December 20, 2025

| | | |
|--------------------------------|---|----------------|
| Research Interests | Design secure and private Machine Learning systems in the presence of untrustworthy | |
| | 1. Providers: Confidential computing via Homomorphic Encryption & Secret Sharing. 2. Data: Mitigate data poisoning during training & prompt injection during inference. 3. Models: Protect training data privacy through PII scrubbing & differential privacy. 4. Users: Control misuse by detecting generated (mis)information with watermarking. | |
| Education | University of Waterloo , Canada | 2019 - 02/2024 |
| | Ph.D. in Computer Science <ul style="list-style-type: none">▪ Advisor: Florian Kerschbaum▪ Thesis: Analyzing Threats of Large-Scale Machine Learning Systems▪ Thesis Awards: Top Mathematics Doctoral Prize & Alumni Gold Medal | |
| | RWTH-Aachen , Germany | 2016 - 2018 |
| | M.Sc. in Computer Science (<i>w/Distinction</i>) | 10/2012 - 2016 |
| | B.Sc. in Computer Science | |
| Honors & Awards | Amazon Research Award [100,000 USD] 2025 First Place at the NeurIPS'24 Watermarking Competition [4,400 USD] 2024 First Place at DGE Elite Hackathon, GITEX'24 [10,900 USD] 2024 Top Mathematics Doctoral Thesis , University of Waterloo [1 080 USD] 2024 Alumni Gold Medal , One PhD Award Yearly, University of Waterloo 2024 Best Poster Award , Sponsored by David R. Cheriton [220 USD] 2023 Distinguished Contribution Award , Microsoft MLADS conference 2023 David R. Cheriton Scholarship , University of Waterloo [14 400 USD] 2022, 2023 Outstanding Reviewer , ICML'22 2022 Best Poster Award , Sponsored by Rogers [720 USD] 2019 | |
| Conference Publications | [NeurIPS'25] Mask Image Watermarking AR: 24.5% (5 290/21 575) Runyi Hu, Jie Zhang, Shiqian Zhao, Nils Lukas , Jiwei Li, Qing Guo, Han Qiu, Tianwei Zhang. The Thirty-Ninth Annual Conference on Neural Information Processing Systems, 2025. | |
| | [EMNLP'25] SPIRIT: Patching Speech Language Models against Jailbreak Attacks AR: 22.2% (1 811/8 172) Amir Djanibekov, Nurgaulet Mukhituly, Kentaro Inui, Hanan Aldarmaki, Nils Lukas . Empirical Methods in Natural Language Processing (Main Conference), 2025. | |
| | [ICML'25] Optimizing Adaptive Attacks against Content Watermarks for Language Models AR: 26.9% (3 260/12 107) Abdulrahman Diaa, Toluwani Aremu, Nils Lukas . The Forty-Second International Conference on Machine Learning, 2025. | |
| | [ICML'25] Cowpox: Towards the Immunity of VLM-based Multi-Agent Systems AR: 26.9% (3 260/12 107) Yutong Wu, Jie Zhang, Yiming Li, Chao Zhang, Qing Guo, Han Qiu, Nils Lukas , Tianwei Zhang. The Forty-Second International Conference on Machine Learning, 2025. | |
| | [USENIX'24] PEPSI: Practically Efficient Private Set Intersection in the Unbalanced Setting AR: 19.1% (417/2176) Rasoul Mahdavi, Nils Lukas , Faezeh Ebrahimianghzani, Thomas Humphries, Bailey Kacsma, John Premkumar, Xinda Li, Simon Oya, Ehsan Amjadian, Florian Kerschbaum. In the 33rd USENIX Security Symposium, 2024. | |

| | |
|--|--|
| [USENIX'24] AR: 19.1% (417/2176) | Fast and Private Inference of Deep Neural Networks by Co-designing Activation Functions Abdulrahman Diaa, Lucas Fenaux, Thomas Humphries, Marian Dietz, Faezeh Ebrahimianghzani, Bailey Kacsma, Xinda Li, Nils Lukas , Rasoul Akhavan Mahdavi, Simon Oya, Ehsan Amjadian, Florian Kerschbaum. In the 33rd USENIX Security Symposium, 2024. |
| [ICLR'24] AR: 30.8% (2250/7262) | Leveraging Optimization for Adaptive Attacks on Image Watermarks Nils Lukas , Abdulrahman Diaa, Lucas Fenaux, Florian Kerschbaum. In the Twelfth International Conference on Learning Representations, 2024. |
| [ICLR'24] AR: 30.8% (2250/7262)  Media Coverage | Universal Backdoor Attacks Benjamin Schneider, Nils Lukas , Florian Kerschbaum. In the Twelfth International Conference on Learning Representations, 2024. |
| [USENIX'23] AR: 29.2% (422/1444) | PTW: Pivotal Tuning Watermarking for Pre-Trained Image Generators Nils Lukas and Florian Kerschbaum. In the 32nd USENIX Security Symposium, 2023. |
| [S&P'23] AR: 17.0% (195/1147)  Distinguished Contribution Award at Microsoft MLADS | Analyzing Leakage of Personally Identifiable Information in Language Models Nils Lukas , Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, Santiago Zanella-Béguelin. In the 44th IEEE Symposium on Security and Privacy, 2023. |
| [S&P'22] AR: 14.5% (147/1012) | SoK: How Robust is Image Classification Deep Neural Network Watermarking? Nils Lukas , Edward Jiang, Xinda Li, Florian Kerschbaum. In the 43rd IEEE Symposium on Security and Privacy, 2022. |
| [ICLR'21] AR: 28.7% (860/2997)  Spotlight (Top 5%) | Deep Neural Network Fingerprinting by Conferrable Adversarial Examples Nils Lukas , Yuxuan Zhang, Florian Kerschbaum. The Ninth International Conference on Learning Representations, 2021. |
| [IH&MMSEC'21] AR: 40.3% (128/318) | On the Robustness of Backdoor-based Watermarking in Deep Neural Networks Masoumeh Shafieinejad, Nils Lukas , Jiaqi Wang, Xinda Li, Florian Kerschbaum. Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, 2021. |
| [ACSAC'20] AR: 20.9% (104/497) | Practical Over-Threshold Multi-Party Private Set Intersection Rasoul Mahdavi, Thomas Humphries, Bailey Kacsma, Simeon Krastnikov, Nils Lukas , John Premkumar, Masoumeh Shafieinejad, Simon Oya, Florian Kerschbaum, Erik-Oliver Blass. Annual Computer Security Applications Conference (ACSAC), 2020. |
| [EuroS&P'20] AR: 20.9% (39/187) | Differentially Private Two-Party Set Operations Bailey Kacsma, Basit Khurram, Nils Lukas , Alexander Norton, Masoumeh Shafieinejad, Zhiwei Shang, Yaser Baseri, Maryam Sepehri, Simon Oya, Florian Kerschbaum. IEEE European Symposium on Security and Privacy (EuroS&P), 2020. |
| Journal Publications | [AIP'18] SunFlower: A new Solar Tower Simulation Method for use in Field Layout Optimization, Pascal Richter, Gregor Heiming, Nils Lukas , Martin Frank. AIP Conference Proceedings, Volume 2033, Issue 1, 2018. |

| | | |
|------------------------|--|---|
| Workshop Papers | [GenAI4Health] | Sanitizing Medical Documents with Differential Privacy using Large Language Models Rushil Thareja, Gautam Gupta, Preslav Nakov, Praneeth Vepakomma, Nils Lukas . 2025. |
| | [WMARK'25] | First-Place Solution to NeurIPS 2024 Invisible Watermark Removal Challenge Fahad Shamshad, Tameem Bakr, Yahia Salaheldin Shaaban, Noor Hazim Hussein, Karthik Nandakumar and Nils Lukas . The 1st Workshop on GenAI Watermarking, 2025. |
| | [WMARK'25]  Oral Presentation | Optimizing Adaptive Attacks against Content Watermarks for Language Models Abdulrahman Diaa, Toluwani Aremu, Nils Lukas . The 1st Workshop on GenAI Watermarking, 2025. |
| Working Papers | | Differentially Private Inference for Large Language Models, Submitted. Rushil Thareja, Preslav Nakov, Praneeth Vepakomma, Nils Lukas . 2025. |
| | | Mitigating Watermark Forgery in Generative Models via Randomized Key Selection, Submitted. Toluwani Aremu, Noor Hazim Hussein, Munachiso S Nwadike, Samuele Poppi, Jie Zhang, Karthik Nandakumar, Neil Zhenqiang Gong, Nils Lukas . 2025. |
| | | Collaborative Threshold Watermarking, Submitted. Tameem Bakr, Anish Ambreth, Nils Lukas . 2025. |
| | | Robust and Calibrated Detection of Authentic Multimedia Content, Submitted. Sarim Hashmi, Abdelrahman Elsayed, Mohammed Talha Alam, Samuele Poppi, Nils Lukas . 2025. |
| Research Talks | Adaptively Robust and Forgery-Resistant Watermarking | 2025 |
| | ▪ Meta (FAIR), hosted by Hady Elsahar | |
| | Optimizing Adaptive Attacks against Content Watermarks | |
| | ▪ DeepMind, hosted by David Stutz | 2024 |
| | ▪ University of California, Berkeley, hosted by Dawn Song | 2024 |
| | Analyzing Leakage of Personal Information in Language Models | |
| | ▪ Microsoft M365, hosted by Robert Sim | 2024 |
| | ▪ Meta, hosted by Will Bullock | 2023 |
| | ▪ MongoDB, hosted by Marilyn George and Archita Agarwal | 2023 |
| | How Reliable is Watermarking for Image Generators? | |
| | ▪ Google, hosted by Somesh Jha | 2023 |
| | ▪ University of California, Berkely, hosted by Dawn Song | 2023 |
| Keynotes | Aviation Future Week , hosted by Emirates, Dubai | 2024 |
| | Cyber Energy Leadership Forum , Abu Dhabi | 2024 |
| Work Experience | Assistant Professor , MBZUAI, Abu Dhabi, UAE | since 08/2024 |
| | Research Intern , Royal Bank of Canada, Borealis AI, Toronto | 2024 |
| | ▪ Vertical Federated Learning, hosted by Kevin Wilson | |
| | Research Intern , Microsoft Research, Cambridge, UK | 2022 |
| | ▪ Privacy for Language Models, hosted by Shruti Tople & Lukas Wutschitz | |
| | Research Assistant , RWTH-Aachen, Aachen | 2014 - 2018 |
| | Student Researcher , DSA Daten- und Systemtechnik GmbH, Aachen | 2016 |
| | Software Engineer Intern , A.R. Bayer DSP Systeme GmbH, Düsseldorf | 2012 |

| | | |
|------------------------|---|--|
| Research Grants | Awarded | 2025 |
| | <ul style="list-style-type: none"> ▪ [Etihad Airways] Conversational Booking Agents. PI: Nils Lukas, Co-PIs: Salem Lahlou, Alham Fikri, Martin Takac, Mingming Gong [450 000 USD] ▪ [United AI-Saqr Group] Privacy-preserving Brain Computer Interfaces. PI: Abdulrahman Mahmoud, Co-PIs: Nils Lukas, Elizabeth Churchill [136 000 USD] ▪ [Amazon Research Awards - Special Call] AdvSim2Real: Robustifying Agents in Simulated Adversarial Environments, PI: Nils Lukas, Co-PI: Praneeth Vepakomma [100 000 USD] | |
| | Selected | |
| | <ul style="list-style-type: none"> ▪ [TII Funding] GFlowNets for Fuzzing of Agentic Applications. PIs: Salem Lahlou & Nils Lukas [136 000 USD] | 2025 |
| Teaching | Instructor , MBZUAI, UAE | |
| | <ul style="list-style-type: none"> ▪ ML8502: Machine Learning Security (14 weeks) ▪ ML807: Federated Learning (7 weeks) ▪ ML818: Emerging Topics in Trustworthy Machine Learning (4 weeks) | <ul style="list-style-type: none"> 2025 2025 2024 |
| | Teaching Assistant , University of Waterloo, Canada | |
| | <ul style="list-style-type: none"> ▪ CS458/658: Computer Security and Privacy ▪ CS246 - Object Oriented Programming | <ul style="list-style-type: none"> 2020, 2021 2021 |
| | Co-Instructor , RWTH-Aachen, Germany | |
| | <ul style="list-style-type: none"> ▪ Course: Data-driven Medicine | 2018 |
| Service | Area Chair | |
| | <ul style="list-style-type: none"> ▪ International Conference on Learning Representations (ICLR) ▪ International Conference on Machine Learning (ICML) | <ul style="list-style-type: none"> 2026 2026 |
| | Program Committee | |
| | <ul style="list-style-type: none"> ▪ ACM Conference on Computer and Communications Security (CCS) ▪ IEEE Symposium on Security and Privacy (IEEE S&P) ▪ Recent Advances in Intrusion Detection (RAID) | <ul style="list-style-type: none"> 2025 2025, 2026 2024 |
| | Artifact Evaluation Committee | |
| | <ul style="list-style-type: none"> ▪ The ACM Conference on Computer and Communications Security (CCS) | 2023, 2024 |
| | Reviewer | |
| | <ul style="list-style-type: none"> ▪ NETYS ▪ ACM TheWebConf (WWW) ▪ International Conference on Learning Representations (ICLR) ▪ International World Wide Web Conference (TheWebConf) ▪ Recent Advances in Intrusion Detection (RAID) ▪ Neural Information Processing Systems (NeurIPS) ▪ International Conference on Machine Learning (ICML) ▪ The Conference on Information and Knowledge Management (CIKM) | <ul style="list-style-type: none"> 2025 2025 2024, 2025 2024 2023 2022, 2023 2022, 2025 2020 |
| | Other | |
| | <ul style="list-style-type: none"> ▪ Sub-Reviewer, Proceedings on Privacy Enhancing Technologies (PETS) ▪ Session Chair, IEEE Symposium on Security and Privacy (S&P) ▪ Organization, Workshop on Semantic Web Solutions for Large-Scale Biomedical Data Analytics (SeWeBMeDA) ▪ Chair for the invited faculty talk program, International Symposium on Trustworthy Foundation Models at MBZUAI ▪ Faculty Search Committee, Machine Learning Department at MBZUAI ▪ Admission's Committee, MBZUAI Machine Learning Department | <ul style="list-style-type: none"> 2021, 2022, 2023 2023 2018 2025 2025 2025 |
| | Student Board Member , Cybersecurity and Privacy Institute | 2022, 2023, 2024 |
| | School Advisory Committee on Appointments Liaison , CrySP Lab | 2022 |

| | | | |
|-------------------------------|------------------|---|---------|
| PostDocs, Students | PostDoc | ▪ Samuele Poppi, Kshitij Mishra, Dushyant Chauhan, Yuhan Liu | 07/2025 |
| | PhD | | |
| | <i>Primary</i> | | |
| | | ▪ Fabian Zhang (ML), Kseniia Kuvshinova (ML) | 10/2025 |
| | | ▪ Toluwani Aremu (ML) | 10/2024 |
| | <i>Secondary</i> | | |
| | | ▪ Tair Djanibekov (NLP), Vamshi Krishna Bonagiri (NLP), Abhijith Sharma (ML) | 10/2025 |
| | | ▪ Rushil Thareja (NLP), Thuan Hoang Nguyen (CV) | 10/2024 |
| | MSc | | |
| | <i>Primary</i> | | |
| | | ▪ Juan Nicolas Sepulveda Arias (CV), Mohamed Mahmoud Mohamed Hendy (ML), Majid Mohamed Sulaiman Ibrahim (ML), Prince Jha (ML), Muhammad Mohideen (CV) | 10/2025 |
| | | ▪ Tameem Bakr (ML), Shaikha Jasem Mohamed Ismail Alhosani (ML), Maryam Abdulla Rashed Abdulla Alshamsi (ML), Maryam Alshehyari (ML), Abdalla Khalid Mohamed Abdalla Alzaabi (ML), Ali Al-Ali (ML) | 10/2024 |
| | <i>Secondary</i> | | |
| | | ▪ Bilal Ashfaq (STATS) | 10/2025 |