December 11, 2023
University of Waterloo, Ontario, Canada
nilslukas.github.io / nlukas@uwaterloo.ca

Research Overview

My research focuses on threats in deploying deep neural networks, encompassing (i) controlling misuse from untrustworthy users, (ii) model integrity with untrustworthy data, and (iii) privacy concerns from untrustworthy models that leak sensitive information. My contributions include studying privacy attacks on language models, defenses against data poisoning, and watermarking methods to control misuse.

Education

Ph.D. in Computer Science, University of Waterloo, Canada — Jan 2019 - Jan 2024
Advisor: Prof. Florian Kerschbaum
Thesis: "Analyzing Threats of Large-Scale Machine Learning Systems"

M.Sc. in Computer Science, RWTH-Aachen, Germany — May 2016 - Oct 2018
Advisor: Prof. Stefan Decker
Thesis: "Secure Inference for Deep Neural Networks"

B.Sc. in Computer Science, RWTH-Aachen, Germany — Oct 2012 - Apr 2016

Experience

Research Intern, Microsoft Research — May 2022 - Aug 2022
Topic: Privacy in Large Language Models. See our IEEE S&P'23 publication.

Instructional Advisor, University of Waterloo — Sep 2021 - Dec 2021
Course: Object-Oriented Software Development

Co-Instructor, Group on Information Systems at the RWTH-Aachen — Oct 2017 - Mar 2018
Course: Data-Driven Medicine

Conference Publications

[USENIX'23]
AR: 29% (422/1 444)

Nils Lukas and Florian Kerschbaum. *PTW: Pivotal Tuning Watermarking for Pre-Trained Image Generators*, 32nd USENIX Security Symposium, 2023, [pdf]

[IEEE S&P'23]
AR: 17% (195/1 147)
🏆 **Distinguished Contribution Award at Microsoft MLADS**

Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, Santiago Zanella-Béguelin. *Analyzing Leakage of Personally Identifiable Information in Language Models*, 44th IEEE Symposium on Security and Privacy, 2023, [pdf]

[IEEE S&P'22]
AR: 14.5% (147/1 012)

Nils Lukas, Edward Jiang, Xinda Li, Florian Kerschbaum. *Sok: How Robust is Image Classification Deep Neural Network Watermarking?*, 43rd IEEE Symposium on Security and Privacy, 2022, [pdf]

[ICLR'21]
AR: 28.7% (860/2 997)
🏆 **Spotlight Award**

Nils Lukas, Yuxuan Zhang, Florian Kerschbaum. *Deep Neural Network Fingerprinting by Conferrable Adversarial Examples*, The Ninth International Conference on Learning Representations, 2021 [pdf]

[IH&MMSEC'21]
AR: 40.0% (128/318)

Masoumeh Shafieinejad, **Nils Lukas**, Jiaqi Wang, Xinda Li, Florian Kerschbaum. *On the Robustness of Backdoor-based Watermarking in Deep Neural Networks*, Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security, 2021, [pdf]

[ACSAC'20]
AR: 20.9% (104/497)

Rasoul Akhavan Mahdavi, Thomas Humphries, Bailey Kacsmar, Simeon Krastnikov, **Nils Lukas**, John A Premkumar, Masoumeh Shafieinejad, Simon Oya, Florian Kerschbaum, Erik-Oliver Blass. *Practical Over-Threshold Multi-Party Private Set Intersection*, Annual Computer Security Applications Conference, 2020, pp. 772-783, [pdf]

[IEEE EuroS&P'20]
AR: 20.9% (39/187)

Bailey Kacsmar, Basit Khurram, **Nils Lukas**, Alexander Norton, Masoumeh Shafieinejad, Zhiwei Shang, Yaser Baseri, Maryam Sepehri, Simon Oya, Florian Kerschbaum. *Differentially private two-party set operations*, 2020 IEEE European Symposium on Security and Privacy (EuroS&P), 2020, pp. 390-404 [pdf]

Journal Publications

[AIP'18] Pascal Richter, Gregor Heiming, **Nils Lukas**, Martin Frank. *SunFlower: A new Solar Tower Simulation Method for use in Field Layout Optimization*, AIP Conference Proceedings, Volume 2033, Issue 1, 2018, [pdf]

Under Submission

2023 **Nils Lukas**, Abdulrahman Diaa, Lucas Fenaux, Florian Kerschbaum. *Leveraging Optimization for Adaptive Attacks on Image Watermarks*, 2023, [pdf]

2023 Benjamin Schneider, **Nils Lukas**, Florian Kerschbaum. *Universal Backdoor Attacks*, 2023, [pdf] [news].
Media Coverage

2023 **Nils Lukas**, Florian Kerschbaum. *Pick your Poison: Undetectability versus Robustness in Data Poisoning Attacks against Deep Image Classification*, 2023, [pdf]

2023 Abdulrahman Diaa, Lucas Fenaux, Thomas Humphries, Marian Dietz, Faezeh Ebrahimianghazani, Bailey Kacsmar, Xinda Li, **Nils Lukas**, Rasoul Akhavan Mahdavi, Simon Oya, Ehsan Amjadian, Florian Kerschbaum. *Fast and Private Inference of Deep Neural Networks by Co-designing Activation Functions*, 2023, [pdf]

2023 Rasoul Mahdavi, **Nils Lukas**, Faezeh Ebrahimianghazani, Thomas Humphries, Bailey Kacsmar, John Premkumar, Xinda Li, Simon Oya, Ehsan Amjadian, Florian Kerschbaum *PEPSI: Practically Efficient Private Set Intersection in the Unbalanced Setting*, 2023.

Tutorials

[ESWC'18] **Nils Lukas**, Oya Beyan, Ali Hasnain. *Privacy-Preserving Information Extraction with Bloom Filters*, European Semantic Web Conference, Greece, 2018, [web]

Awards and Honors

2023 **Best Poster Award** - Presentation with David R. Cheriton [300 CAD]
Distinguished Contribution Award - Microsoft-internal 2023 MLADS conference
2022 **David R. Cheriton Scholarship** - 2-year award [20,000 CAD]
Outstanding Reviewer - International Conference on Machine Learning (ICML'22)
2019 **Best Poster Award** - Cybersecurity and Privacy Institute (CPI) [1,000 CAD]
2018 **Excellence Graduation** - Master's degree
2016 **KU Global Scholarship** - Semester abroad, Korea University [1.2 million KRW]
2014 **MOGAM Scholarship** - Semester abroad, Yonsei University, Korea [3,000 EUR]

University Service

2022 **Student Board Member** - Cybersecurity and Privacy Institute (CPI)
School Advisory Committee on Appointments (SACA) Liaison - CrySP lab

Academic Service

Program Chair Artifact Evaluation for CCS'23
Session Chair IEEE S&P'23
External Reviewer TheWebConf'24, ICLR'24, RAID'23, PETS'23, NeurIPS'23, ICML'22 🏆, PETS'22, NeurIPS'22, PETS'21, CIKM'20

Teaching

2018 *Data-Driven Medicine*, RWTH-Aachen

Invited Talks

Dec, 2023	*Analyzing Leakage of Personally Identifiable Information in Language Models*, Privacy-Preserving Machine Learning Group at Meta, [slides]
Oct, 2023	*How Reliable is Watermarking for Generative Machine Learning?*, University of California, Berkeley, Prof. Dawn Song's Group, [slides]
June, 2023	*A Learnable Watermark for Deep Image Generators*, Google's Red Team [slides]
June, 2023	*Analyzing Leakage of Personally Identifiable Information in Language Models*, MongoDB, [slides]