Nicolas Papernot

10 King's College Road – Toronto, Ontario M5S 3G4 – Canada

✉ nicolas.papernot@utoronto.ca • 🌐 www.papernot.fr

Academic and Research Appointments

University of Toronto **Toronto, ON, Canada**
Assistant Professor *Since 09/2019*

- *Department of Electrical & Computer Engineering*
- *Department of Computer Science*

Vector Institute **Toronto, ON, Canada**
Canada CIFAR AI Chair and Faculty Member *Since 09/2019*

Google DeepMind **Mountain View, CA, USA**
Research Scientist (on the Brain team) *Since 08/2018*

Google Brain **Mountain View, CA, USA**
Research Intern (mentored by Ilya Mironov) *05/2017–12/2017*

Google Research **Mountain View, CA, USA**
Research Intern (mentored by Ulfar Erlingsson and Martin Abadi) *05/2016–08/2016*

Education

Pennsylvania State University **University Park, PA, USA**
Ph.D. in Computer Science and Engineering *2016–2018*

- Dissertation: *Characterizing the Limits and Defenses of Machine Learning in Adversarial Settings*
- Advisor: Prof. Patrick McDaniel
- Dissertation committee: Prof. Patrick McDaniel, Prof. Trent Jaeger, Prof. Thomas F. LaPorta, Prof. Aleksandra Slavkovic, Prof. Dan Boneh, Dr. Ian J. Goodfellow

Pennsylvania State University **University Park, PA, USA**
M.S. in Computer Science and Engineering *2014–2016*

- Thesis: *On The Integrity Of Deep Learning Systems in Adversarial Settings*
- Advisor: Prof. Patrick McDaniel
- Thesis committee: Prof. Patrick McDaniel, Prof. Adam D. Smith

École Centrale de Lyon **Lyon, France**
Diplôme d'Ingénieur (M.S. and B.S. in Engineering Sciences) *2012–2016*

Lycée Louis-le-Grand **Paris, France**
Classe Préparatoire (equivalent to first two years of B.S. in the US and Canada) *2010–2012*

Honors

AI2050 Early Career Fellowship: Schmidt Sciences *2024*
Member of the College of New Scholars: Royal Society of Canada *2023*
Outstanding Paper Award: 10th International Conference on Learning Representations *2022*

Alfred P. Sloan Research Fellow in Computer Science: Sloan Foundation	*2022*
Early Researcher Award: Ministry of Colleges and Universities	*2022*
Outstanding Performance Discretionary Research Grant: Vector Institute	*2021*
Faculty Affiliate: Schwartz Reisman Institute	*2020-2024*
Connaught New Researcher Award: University of Toronto	*2020*
Canada CIFAR AI Chair: Canadian Institute for Advanced Research	*2019*
Top 30% Reviewers Award: Neural Information Processing Systems	*2018*
Wormley Family Graduate Fellowship: Pennsylvania State University	*2018*
CSE Research Assistant Award: Pennsylvania State University	*2018*
Student Travel Award: 6th International Conference on Learning Representations	*2018*
Student Travel Award: 34th International Conference on Machine Learning	*2017*
Best Paper Award: 5th International Conference on Learning Representations	*2017*
Student Travel Award: 5th International Conference on Learning Representations	*2017*
CSE Graduate Research Award: Pennsylvania State University	*2016*
Google PhD Fellowship in Security: Google Research	*2016–2018*
CyberSpace 2025 Essay Contest (2nd place): Microsoft	*2015*
Scholarship for Exceptional Academic Achievements: McGill (declined)	*2010*

Publications

Selected Pre-prints

Unlearnable Algorithms for In-context Learning. *Andrei Ioan Muresanu, Anvith Thudi, Michael R. Zhang, Nicolas Papernot*. (2024)

Regulation Games for Trustworthy Machine Learning. *Mohammad Yaghini, Patty Liu, Franziska Boenisch, Nicolas Papernot*. (2024)

LLM Censorship: A Machine Learning Challenge or a Computer Security Problem?. *David Glukhov, Ilia Shumailov, Yarin Gal, Nicolas Papernot, Vardan Papyan*. (2023)

The Curse of Recursion: Training on Generated Data Makes Models Forget. *Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, Ross Anderson*. (2023)

Gradients Look Alike: Sensitivity is Often Overestimated in DP-SGD. *Anvith Thudi, Hengrui Jia, Casey Meehan, Ilia Shumailov, Nicolas Papernot*. (2023)

Selective Classification Via Neural Network Training Dynamics. *Stephan Rabanser, Anvith Thudi, Kimia Hamidieh, Adam Dziedzic, Nicolas Papernot*. (2022)

Conference proceedings

Memorization in Self-Supervised Learning Improves Downstream Generalization. *Wenhao Wang, Muhammad Ahmad Kaleem, Adam Dziedzic, Michael Backes, Nicolas Papernot, Franziska Boenisch*. Proceedings of the 12th International Conference on Learning Representations. (2024)

Confidential-DPproof: Confidential Proof of Differentially Private Training. *Ali Shahin Shamsabadi, Gefei Tan, Tudor Ioan Cebere, Aurélien Bellet, Hamed Haddadi, Nicolas Papernot, Xiao Wang, Adrian Weller*. Proceedings of the 12th International Conference on Learning Representations. **[Spotlight Paper Award]** (2024)

Exploring Strategies for Guiding Symbolic Analysis with Machine Learning Prediction. *Mingyue Yang, David Lie, Nicolas Papernot*. 31st IEEE International Conference on Software Analysis, Evolution and Reengineering. (2024)

Robust and Actively Secure Serverless Collaborative Learning. *Olive Franzese, Adam Dziedzic, Christopher A. Choquette-Choo, Mark R. Thomas, Muhammad Ahmad Kaleem, Stephan Rabanser, Congyu Fang, Somesh Jha, Nicolas Papernot, Xiao Wang*. Proceedings of the 37th Conference on Neural Information Processing Systems. (2023)

Training Private Models That Know What They Don't Know. *Stephan Rabanser, Anvith Thudi, Abhradeep Thakurta, Krishnamurthy Dvijotham, Nicolas Papernot*. Proceedings of the 37th Conference on Neural Information Processing Systems. (2023)

Have it your way: Individualized Privacy Assignment for DP-SGD. *Franziska Boenisch, Christopher Mühl, Adam Dziedzic, Roy Rinberg, Nicolas Papernot*. Proceedings of the 37th Conference on Neural Information Processing Systems. (2023)

Flocks of Stochastic Parrots: Differentially Private Prompt Learning for Large Language Models. *Haonan Duan, Adam Dziedzic, Nicolas Papernot, Franziska Boenisch*. Proceedings of the 37th Conference on Neural Information Processing Systems. (2023)

Proof-of-Learning is Currently More Broken Than You Think. *Congyu Fang, Hengrui Jia, Anvith Thudi, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Varun Chandrasekaran, Nicolas Papernot*. Proceedings of the 8th IEEE European Symposium on Security and Privacy, Delft, Netherlands. (2023)

Reconstructing Individual Data Points in Federated Learning Hardened with Differential Privacy and Secure Aggregation. *Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, Nicolas Papernot*. Proceedings of the 8th IEEE European Symposium on Security and Privacy, Delft, Netherlands. (2023)

When the Curious Abandon Honesty: Federated Learning Is Not Private. *Franziska Boenisch, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, Nicolas Papernot*. Proceedings of the 8th IEEE European Symposium on Security and Privacy, Delft, Netherlands. (2023)

Losing Less: A Loss for Differentially Private Deep Learning. *Ali Shahin Shamsabadi, Nicolas Papernot*. Proceedings on Privacy Enhancing Technologies, Lausanne, Switzerland. (2023)

Architectural Backdoors in Neural Networks. *Mikel Bober-Irizar, Ilia Shumailov, Yiren Zhao, Robert Mullins, Nicolas Papernot*. Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada. (2023)

Measuring Forgetting of Memorized Training Examples. *Matthew Jagielski, Om Thakkar, Florian Tramer, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, Chiyuan Zhang*. Proceedings of the 11th International Conference on Learning Representations. (2023)

Confidential-PROFITT: Confidential PROof of FaIr Training of Trees. *Ali Shahin Shamsabadi, Sierra Calanda Wyllie, Nicholas Franzese, Natalie Dullerud, Sébastien Gambs, Nicolas Papernot, Xiao Wang, Adrian Weller*. Proceedings of the 11th International Conference on Learning Representations. **[Oral Paper Award]** (2023)

Private Multi-Winner Voting for Machine Learning. *Adam Dziedzic, Christopher A. Choquette-Choo, Natalie Dullerud, Vinith Menon Suriyakumar, Ali Shahin Shamsabadi, Muhammad Ahmad Kaleem, Somesh Jha, Nicolas Papernot, Xiao Wang*. Proceedings on Privacy Enhancing Technologies, Lausanne, Switzerland. (2023)

Differentially Private Speaker Anonymization. *Ali Shahin Shamsabadi, Brij Mohan Lal Srivastava, Aurelien Bellet, Nathalie Vauquier, Emmanuel Vincent, Mohamed Maouche, Marc Tommasi, Nicolas Papernot*. Proceedings on Privacy Enhancing Technologies, Lausanne, Switzerland. (2023)

Tubes Among Us: Analog Attack on Automatic Speaker Identification. *Shimaa Ahmed, Yash Wani, Ali Shahin Shamsabadi, Mohammad Yaghini, Ilia Shumailov, Nicolas Papernot, Kassem Fawaz*. Proceedings of the 32nd USENIX Security Symposium. (2023)

The Privacy Onion Effect: Memorization is Relative. *Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, Florian Tramer*. Proceedings of the 36th Conference on Neural Information

Processing Systems. (2022)

Washing The Unwashable: On The (Im)possibility of Fairwashing Detection. *Ali Shahin Shamsabadi, Mohammad Yaghini, Natalie Dullerud, Sierra Wyllie, Ulrich Aïvodji, Aisha Alaagib Alryeh Mkean, Sébastien Gambs, Nicolas Papernot*. Proceedings of the 36th Conference on Neural Information Processing Systems. (2022)

Dataset Inference for Self-Supervised Models. *Adam Dziedzic, Haonan Duan, Muhammad Ahmad Kaleem, Nikita Dhawan, Jonas Guan, Yannis Cattan, Franziska Boenisch, Nicolas Papernot*. Proceedings of the 36th Conference on Neural Information Processing Systems. (2022)

In Differential Privacy, There is Truth: on Vote-Histogram Leakage in Ensemble Private Learning. *Jiaqi Wang, Roei Schuster, Ilia Shumailov, David Lie, Nicolas Papernot*. Proceedings of the 36th Conference on Neural Information Processing Systems. (2022)

On the Limitations of Stochastic Pre-processing Defenses. *Yue Gao, Ilia Shumailov, Kassem Fawaz, Nicolas Papernot*. Proceedings of the 36th Conference on Neural Information Processing Systems. (2022)

On the Difficulty of Defending Self-Supervised Learning against Model Extraction. *Adam Dziedzic, Nikita Dhawan, Muhammad Ahmad Kaleem, Jonas Guan, Nicolas Papernot*. Proceedings of the 39th International Conference on Machine Learning. (2022)

Unrolling SGD: Understanding Factors Influencing Machine Unlearning. *Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, Nicolas Papernot*. Proceedings of the 7th IEEE European Symposium on Security and Privacy, Genoa, Italy. (2022)

On the Necessity of Auditable Algorithmic Definitions for Machine Unlearning. *Anvith Thudi, Hengrui Jia, Ilia Shumailov, Nicolas Papernot*. Proceedings of the 31st USENIX Security Symposium. (2022)

Increasing the Cost of Model Extraction with Calibrated Proof of Work. *Adam Dziedzic, Muhammad Ahmad Kaleem, Yu Shen Lu, Nicolas Papernot*. Proceedings of the 10th International Conference on Learning Representations. **[Spotlight Paper Award]** (2022)

A Zest of LIME: Towards Architecture-Independent Model Distances. *Hengrui Jia, Hongyu Chen, Jonas Guan, Ali Shahin Shamsabadi, Nicolas Papernot*. Proceedings of the 10th International Conference on Learning Representations. (2022)

Hyperparameter Tuning with Renyi Differential Privacy. *Nicolas Papernot, Thomas Steinke*. Proceedings of the 10th International Conference on Learning Representations. **[Outstanding Paper Award]** (2022)

Is Fairness Only Metric Deep? Evaluating and Addressing Subgroup Gaps in Deep Metric Learning. *Natalie Dullerud, Karsten Roth, Kimia Hamidieh, Nicolas Papernot, Marzyeh Ghassemi*. Proceedings of the 10th International Conference on Learning Representations. (2022)

Bad Character Injection: Imperceptible Attacks on NLP Models. *Nicholas Boucher, Ilia Shumailov, Ross Anderson, Nicolas Papernot*. Proceedings of the 43rd IEEE Symposium on Security and Privacy, San Francisco, CA. (2022)

Towards More Robust Keyword Spotting for Voice Assistants. *Shimaa Ahmed, Ilia Shumailov, Nicolas Papernot, Kassem Fawaz*. Proceedings of the 31st USENIX Security Symposium. (2022)

Manipulating SGD with Data Ordering Attacks. *Ilia Shumailov, Zakhar Shumaylov, Dmitry Kazhdan, Yiren Zhao, Nicolas Papernot, Murat A. Erdogdu, Ross Anderson*. Proceedings of the 35th Conference on Neural Information Processing Systems. (2021)

Markpainting: Adversarial Machine Learning meets Inpainting. *David Khachaturov, Ilia Shumailov, Yiren Zhao, Nicolas Papernot, Ross Anderson*. Proceedings of the 38th International Conference on Machine Learning. (2021)

Label-Only Membership Inference Attacks. *Christopher A. Choquette-Choo, Florian Tramer, Nicholas Carlini, Nicolas Papernot*. Proceedings of the 38th International Conference on Machine Learning. (2021)

Data-Free Model Extraction. *Jean-Baptiste Truong, Pratyush Maini, Robert Walls, Nicolas Papernot*. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN. (2021)

Proof-of-Learning: Definitions and Practice. *Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, Nicolas Papernot*. Proceedings of the 42nd IEEE Symposium on Security and Privacy, San Francisco, CA. (2021)

Entangled Watermarks as a Defense against Model Extraction. *Hengrui Jia, Christopher A. Choquette-Choo, Varun Chandrasekaran, Nicolas Papernot*. Proceedings of the 30th USENIX Security Symposium. (2021)

Sponge Examples: Energy-Latency Attacks on Neural Networks. *Ilia Shumailov, Yiren Zhao, Daniel Bates, Nicolas Papernot, Robert Mullins, Ross Anderson*. Proceedings of the 6th IEEE European Symposium on Security and Privacy, Vienna, Austria. (2021)

CaPC Learning: Confidential and Private Collaborative Learning. *Christopher A. Choquette-Choo, Natalie Dullerud, Adam Dziedzic, Yunxiang Zhang, Somesh Jha, Nicolas Papernot, Xiao Wang*. Proceedings of the 9th International Conference on Learning Representations. (2021)

Dataset Inference: Ownership Resolution in Machine Learning. *Pratyush Maini, Mohammad Yaghini, Nicolas Papernot*. Proceedings of the 9th International Conference on Learning Representations. **[Spotlight Paper Award]** (2021)

Chasing Your Long Tails: Differentially Private Prediction in Health Care Settings. *Vinith Suriyakumar, Nicolas Papernot, Anna Goldenberg, Marzyeh Ghassemi*. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. (2021)

Adversary Instantiation: Lower bounds for differentially private machine learning. *Milad Nasr, Shuang Song, Abhradeep Guha Thakurta, Nicolas Papernot, Nicholas Carlini*. Proceedings of the 42nd IEEE Symposium on Security and Privacy, San Francisco, CA. (2021)

Tempered Sigmoids for Deep Learning with Differential Privacy. *Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, Ulfar Erlingsson*. Proceedings of the 35th AAAI Conference on Artificial Intelligence. (2021)

Neighbors From Hell: Voltage Attacks Against Deep Learning Accelerators on Multi-Tenant FPGAs. *Andrew Boutros, Mathew Hall, Nicolas Papernot, Vaughn Betz*. Proceedings of the 2020 International Conference on Field-Programmable Technology. (2020)

Machine Unlearning. *Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, Nicolas Papernot*. Proceedings of the 42nd IEEE Symposium on Security and Privacy, San Francisco, CA. (2021)

SoK: The Faults in our ASRs: An Overview of Attacks against Automatic Speech Recognition and Speaker Identification Systems. *Hadi Abdullah, Kevin Warren, Vincent Bindschaedler, Nicolas Papernot, Patrick Traynor*. Proceedings of the 42nd IEEE Symposium on Security and Privacy, San Francisco, CA. (2021)

Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations. *Florian Tramer, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, Jorn-Henrik Jacobsen*. Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria. (2020)

Thieves of Sesame Street: Model Extraction on BERT-based APIs. *Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, Mohit Iyyer*. Proceedings of the 8th International Conference on Learning Representations, Addis Ababa, Ethiopia. (2020)

High Accuracy and High Fidelity Extraction of Neural Networks. *Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, Nicolas Papernot*. Proceedings of the 29th USENIX Security Symposium. Boston, MA. (2020)

MixMatch: A Holistic Approach to Semi-Supervised Learning. *David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, Colin Raffel*. Proceedings of the 33rd Conference on Neural Information Processing Systems, Vancouver, Canada. (2019)

Analyzing and Improving Representations with the Soft Nearest Neighbor Loss. *Nicholas Frosst, Nicolas Papernot, Geoffrey Hinton*. Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA. (2019)

Adversarial Examples Influence Human Visual Perception. *Gamaleldin F. Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, Jascha Sohl-Dickstein*. Proceedings of the 2019 Computational and Systems Neuroscience meeting, Lisbon, Portugal. (2019)

Adversarial Examples that Fool both Computer Vision and Time-Limited Humans. *Gamaleldin F. Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, Jascha Sohl-Dickstein*. Proceedings of the 32nd Conference on Neural Information Processing Systems, Montreal, Canada. (2018)

Scalable Private Learning with PATE. *Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, Ulfar Erlingsson*. Proceedings of the 6th International Conference on Learning Representations, Vancouver, Canada. (2018)

Ensemble Adversarial Training: Attacks and Defenses. *Florian Tramer, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, Patrick McDaniel*. Proceedings of the 6th International Conference on Learning Representations, Vancouver, Canada. (2018)

Towards the Science of Security and Privacy in Machine Learning. *Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman*. Proceedings of the 3rd IEEE European Symposium on Security and Privacy, London, UK. (2018)

Adversarial Examples for Malware Detection. *Kathrin Grosse, Nicolas Papernot, Praveen Manoharan, Michael Backes, and Patrick McDaniel*. Proceedings of the 2017 European Symposium on Research in Computer Security, Oslo, Norway. (2017)

Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. *Nicolas Papernot, Martin Abadi, Ulfar Erlingsson, Ian Goodfellow, and Kunal Talwar*. Proceedings of the 5th International Conference on Learning Representations, Toulon, France. **[Best Paper Award]** (2017)

Practical Black-Box Attacks against Machine Learning. *Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z.Berkay Celik, and Ananthram Swami*. Proceedings of the 2017 ACM Asia Conference on Computer and Communications Security, Abu Dhabi, UAE. (2017)

Crafting Adversarial Input Sequences for Recurrent Neural Networks. *Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang*. Proceedings of the 2016 Military Communications Conference (MILCOM), Baltimore, MD. (2016)

Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks. *Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami*. Proceedings of the 37th IEEE Symposium on Security and Privacy, San Jose, CA. (2016)

The Limitations of Deep Learning in Adversarial Settings. *Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami*. Proceedings of the 1st IEEE European Symposium on Security and Privacy, Saarbrucken, Germany. (2016)

Enforcing Agile Access Control Policies in Relational Databases using Views. *Nicolas Papernot, Patrick McDaniel, and Robert Walls*. Proceedings of the 2015 Military Communications Conference (MILCOM), Tampa, FL. (2015)

Journals

From Differential Privacy to Bounds on Membership Inference: Less can be More. *Anvith Thudi, Ilia Shumailov, Franziska Boenisch, Nicolas Papernot*. Transactions on Machine Learning Research. (2024)

Decentralised, Collaborative, and Privacy-preserving Machine Learning for Multi-Hospital Data. *Congyu Fang, Adam Dziedzic, Lin Zhang, Laura Oliva, Amol Verma, Fahad Razak, Nicolas Papernot, Bo Wang*. eBioMedicine Volume 101. (2024)

Advancing Differential Privacy: Where are we now and future directions for real-world deployment. *Rachel Cummings, Damien Desfontaines, David Evans, Roxana Geambasu, Yangsibo Huang, Matthew Jagielski, Peter Kairouz, Gautam Kamath, Sewoong Oh, Olga Ohrimenko, Nicolas Papernot, Ryan Rogers, Milan Shen, Shuang Song, Weijie Su, Andreas Terzis, Abhradeep Thakurta, Sergei Vassilvitskii, Yu-Xiang Wang, Li Xiong, Sergey Yekhanin, Da Yu, Huanyu Zhang, Wanrong Zhang*. Harvard Data Science Review. (2024)

Subtle adversarial image manipulations influence both human and machine perception. *Vijay Veerabadran, Josh Goldman, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, Jonathon Shlens, Jascha Sohl-Dickstein, Michael C. Mozer, Gamaleldin F. Elsayed*. Nature Communications. (2022)

Adversarial Examples for Network Intrusion Detection Systems. *Ryan Sheatsley and Nicolas Papernot and Michael J. Weisman and Gunjan Verma and Patrick McDaniel*. Journal of Computer Security. (2022)

Book Chapters

Differential Privacy and Medical Data Analysis. *Vinith M. Suriyakumar, Nicolas Papernot, Anna Goldenberg, Marzyeh Ghassemi*. Differential Privacy in Artificial Intelligence: From Theory to Practice. (2024)

Private Deep Learning. *Nicolas Papernot*. Differential Privacy in Artificial Intelligence: From Theory to Practice. (2024)

Adversarial Machine Learning. *Nicolas Papernot*. Encyclopedia of Cryptography, Security and Privacy. (2021)

Invited publications

How Relevant Is the Turing Test in the Age of Sophisbots?. *Dan Boneh, Andrew J. Grotto, Patrick McDaniel, Nicolas Papernot*. IEEE Security and Privacy Magazine. (2019)

A Marauder's Map of Security and Privacy in Machine Learning: An overview of current and future research directions for making machine learning secure and private. *Nicolas Papernot*. Keynote at the 11th ACM Workshop on Artificial Intelligence and Security colocated with the 25th ACM Conference on Computer and Communications Security, Toronto, Canada. (2018)

Making Machine Learning Robust against Adversarial Inputs. *Ian Goodfellow, Patrick McDaniel, Nicolas Papernot*. Communications of the ACM. (2018)

On the Protection of Private Information in Machine Learning Systems: Two Recent Approaches. *Martin Abadi, Ulfar Erlingsson, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Nicolas Papernot, Kunal Talwar, Li Zhang*. Proceedings of the 30th IEEE Computer Security Foundations Symposium, Santa Barbara, CA, USA. (2017)

Machine Learning in Adversarial Settings. *Patrick McDaniel, Nicolas Papernot, Z. Berkay Celik*. IEEE Security and Privacy Magazine . (2016)

Policy Briefings

Why We Should Regulate Information About Persons. *Lisa Austin, David Lie, Nicolas Papernot, Aleksandar Nikolov*. Privacy Law Scholars Conference. (2021)

Commentary on Data and Algorithm Privacy. *Aleksandar Nikolov, Nicolas Papernot*. Government of Canada Consultation on the Privacy Act. (2021)

Preparing for the Age of Deepfakes and Disinformation. *Dan Boneh, Andrew J. Grotto, Patrick McDaniel, Nicolas Papernot*. Stanford HAI Policy Brief. (2020)

Workshop publications

Preempt: Sanitizing Sensitive Prompts for LLMs. *Amrita Roy Chowdhury, David Glukhov, Divyam Anshumaan, Prasad Chalasani, Nicolas Papernot, Somesh Jha*. Fifth AAAI Workshop on Privacy-Preserving Artificial Intelligence. (2024)

Regulation Games for Trustworthy Machine Learning. *Mohammad Yaghini, Patty Liu, Franziska Boenisch, Nicolas Papernot*. NeurIPS 2023 Workshop on Regulatable ML. (2023)

Learning with Impartiality to Walk on the Pareto Frontier of Fairness, Privacy, and Utility. *Mohammad Yaghini, Patty Liu, Franziska Boenisch and Nicolas Papernot*. NeurIPS 2023 Workshop on Regulatable ML. (2023)

The Adversarial Implications of Variable-Time Inference. *Dudi Biton, Aditi Misra, Efrat Levy, Jaidip Kotak, Ron Bitton, Roei Schuster, Nicolas Papernot, Yuval Elovici, Ben Nassi*. 16th ACM Workshop on Artificial Intelligence and Security. (2023)

Why is it Gaussian? Exploring the Generalized Gaussian Mechanism for Private Machine Learning. *Roy Rinberg, Ilia Shumailov, Rachel Cummings, Nicolas Papernot*. Theory and Practice of Differential Privacy. (2023)

On the Privacy Risk of In-context Learning. *Haonan Duan, Adam Dziedzic, Mohammad Yaghini, Nicolas Papernot, Franziska Boenisch*. ACL 2023 Workshop on Trustworthy Natural Language Processing. (2023)

Transforming Genomic Interpretability: A DNABERT Case Study. *Micaela Consens, Alan Moses, Bo Wang, Nicolas Papernot*. ICML 2023 Workshop on Computational Biology. (2023)

Sentence Embedding Encoders are Easy to Steal but Hard to Defend. *Adam Dziedzic, Franziska Boenisch, Haonan Duan, Mingjian Jiang, Nicolas Papernot*. ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML. (2023)

Accelerating Symbolic Analysis for Android Apps. *Mingyue Yang, David Lie, Nicolas Papernot*. 4th International Workshop on Advances in Mobile App Analysis. (2021)

Dataset Inference: Ownership Resolution in Machine Learning. *Pratyush Maini, Mohammad Yaghini, Nicolas Papernot*. NeurIPS 2020 workshop on Privacy-preserving Machine Learning. (2020)

Tempered Sigmoids for Deep Learning with Differential Privacy. *Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, Ulfar Erlingsson*. Theory and Practice of Differential Privacy. (2020)

The Pitfalls of Differentially Private Prediction in Healthcare. *Vinith Suriyakumar, Nicolas Papernot, Anna Goldenberg and Marzyeh Ghassemi*. Theory and Practice of Differential Privacy. (2020)

On the Robustness of Cooperative Multi-Agent Reinforcement Learning. *Jieyu Lin, Kristina Dzeparoska, Sai Qian Zhang, Alberto Leon-Garcia, Nicolas Papernot*. Proceedings of the 3rd Deep Learning and Security workshop colocated with the 41st IEEE Symposium on Security and Privacy. (2020)

Improving Differentially Private Models via Active Learning. *Zhengli Zhao, Nicolas Papernot, Sameer Singh, Neoklis Polyzotis, and Augustus Odena*. Presented at the NeurIPS 2019 Workshop on Privacy in Machine Learning. (2019)

Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness. *Jorn-Henrik Jacobsen, Jens Behrmannn, Nicholas Carlini, Florian Tramer, Nicolas Papernot*. Presented at the ICLR 2019 workshop on Safe ML, New Orleans, Louisiana. (2019)

A General Approach to Adding Differential Privacy to Iterative Training Procedures. *Brendan McMahan, Galen Andrew, Ulfar Erlingsson, Steve Chien, Ilya Mironov, Nicolas Papernot, Peter Kairouz*. Presented at the NeurIPS 2018 workshop on Privacy Preserving Machine Learning, Montreal, Canada. (2019)

Extending Defensive Distillation. *Nicolas Papernot and Patrick McDaniel*. Presented at the Workshop track of the 38th IEEE Symposium on Security and Privacy, San Jose, CA. (2017)

Adversarial Attacks on Neural Network Policies. *Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, Pieter Abbeel*. Presented at the Workshop Track of the 5th International Conference on Learning Representations, Toulon, France. (2017)

Security and Science of Agility. *Patrick McDaniel, T. Jaeger, T. F. La Porta, Nicolas Papernot, R. J. Walls, A. Kott, L. Marvel, A. Swami, P. Mohapatra, S. V. Krishnamurthy, I. Neamtiu*. Presented at the 2014 ACM Workshop on

Moving Target Defense. (2014)

Technical reports

Learned Systems Security. *Roei Schuster, Nicolas Papernot, Paul Grubbs, Jin Peng Zhou*. (2022)

Intrinsic Anomaly Detection for Multi-Variate Time Series. *Stephan Rabanser, Tim Januschowski, Kashif Rasul, Oliver Borchert, Richard Kurle, Jan Gasthaus, Michael Bohlke-Schneider, Nicolas Papernot, Valentin Flunkert*. (2022)

Interpretability in Safety-Critical Financial Trading Systems. *Gabriel Deza, Adelin Travers, Colin Rowat, Nicolas Papernot*. (2021)

On the Exploitability of Audio Machine Learning Pipelines to Surreptitious Adversarial Examples. *Adelin Travers, Lorna Licollari, Guanghan Wang, Varun Chandrasekaran, Adam Dziedzic, David Lie, Nicolas Papernot*. (2021)

p-DkNN: Out-of-Distribution Detection through Statistical Testing of Deep Representation. *Adam Dziedzic, Stephan Rabanser, Mohammad Yaghini, Armin Ale, Murat A Erdogdu, Nicolas Papernot*. (2022)

Generative Extraction of Audio Classifiers for Speaker Identification. *Tejumade Afonja, Lucas Bourtoule, Varun Chandrasekaran, Sageev Oore, Nicolas Papernot*. (2022)

On Attribution of Deepfakes. *Baiwu Zhang, Jin Zhou, Ilia Shumailov, Nicolas Papernot*. (2020)

On the Effectiveness of Mitigating Data Poisoning Attacks with Gradient Shaping. *Sanghyun Hong, Varun Chandrasekaran, Yigitcan Kaya, Tudor Dumitras, Nicolas Papernot*. (2020)

Rearchitecting Classification Frameworks For Increased Robustness. *Varun Chandrasekaran, Brian Tang, Nicolas Papernot, Kassem Fawaz, Somesh Jha, Xi Wu*. (2019)

On Evaluating Adversarial Robustness. *Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry*. (2019)

Distribution Density, Tails, and Outliers in Machine Learning: Metrics and Applications. *Nicholas Carlini, Ulfar Erlingsson, Nicolas Papernot*. (2019)

CleverHans v2.1.0: an adversarial machine learning library. *Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin et al.*. (2018)

Deep k-Nearest Neighbors: Towards Confident, Interpretable and Robust Deep Learning. *Nicolas Papernot and Patrick McDaniel*. (2018)

The Space of Transferable Adversarial Examples. *Florian Tramer, Nicolas Papernot, Ian Goodfellow, Dan Boneh, Patrick McDaniel*. (2017)

On the (Statistical) Detection of Adversarial Examples. *Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel*. (2017)

Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow*. (2016)

Dissertation and Thesis

Characterizing the Limits and Defenses of Machine Learning in Adversarial Settings. *Nicolas Papernot*. (2018)

On The Integrity Of Deep Learning Systems In Adversarial Settings. *Nicolas Papernot*. (2016)

Students and postdoctoral fellows

Current group

David Glukhov (co-advised with Vardan Papyan): started Winter 2023 *PhD student*

Mark Thomas [Vector Scholarship in AI, OGS Scholar]: starting Fall 2024 *MSc student*

Ziwen Han: Fall 2023 - Summer 2024 *Undergraduate student*

Emmy Fang (co-advised with Bo Wang) [OGS Scholar, DiDi Award]: started Fall 2023 *PhD student*

Pascale Gourdeau: started Fall 2023, co-advised with Shai Ben-David *Postdoctoral Fellow*

Andy Liu: Fall 2023 - Summer 2024 *Engineering Science student*

Haonan Duan (co-advised with Chris Maddison): started Fall 2021 *PhD student*

Anvith Thudi (co-advised with Chris Maddison) [Vanier Scholar]: started Fall 2022 *PhD student*

Aditi Misra: started Fall 2021 *Engineering Science student*

Sierra Wyllie: started Summer 2021 *Engineering Science student*

Muhammad Ahmad Kaleem: started Summer 2021 *Engineering Science student*

Mohammad Yaghini [Meta PhD Fellow]: started Fall 2020 *PhD student*

Stephan Rabanser: started Fall 2020 *PhD student*

Jonas Guan: started Fall 2020 *PhD student*

Jiaqi Wang (co-advised with David Lie) [OGS Scholar]: started Fall 2020 *MASc student*

Nick Jia [Vector Scholarship, Mary H. Beatty Fellow, OGS Scholar]: started Fall 2020 *PhD student*

Mingyue Yang (co-advised with David Lie): started Winter 2020 *PhD student*

Alumni

Tudor Cebere **Visiting PhD student**
Currently PhD student at INRIA *Fall 2023 - Winter 2024*

Berivan Isik **Visiting PhD student**
Currently PhD student at Stanford University *Summer 2023*

Karan Chadha (co-hosted with Matthew Jagielski) **Google Brain Intern**
Currently PhD student at Stanford University *Summer 2023*

Camille Bruckmann **Engineering Science student**
Currently Software Engineer at Microsoft *Fall 2022 - Summer 2023*

Si Cheng (Steven) Zhong **Engineering Science student**
Currently MS student at University of Toronto *Fall 2022 - Summer 2023*

Franziska Boenisch **Postdoctoral Fellow**
Currently Assistant Professor at CISPA Helmholtz Center for Information Security *Fall 2022 - Summer 2023*

David Glukhov (co-advised with Vardan Papyan) [OGS Scholar] **MS student**
Currently PhD student at University of Toronto *Fall 2022 - Winter 2023*

Shimaa Ahmed **Visiting PhD student**
Currently PhD student at University of Wisconsin-Madison *Summer 2022*

Roy Rinberg **Research Intern**
Currently PhD student at Harvard University *Summer 2022*

Patty Liu **Research Intern**
Currently Undergraduate student at the University of Toronto *May 2022 - August 2023*

Mark Thomas **Research Intern**
Currently MS student at the University of Toronto *Summer 2022*

Avital Shafran **Visiting PhD student**
Currently PhD student at the Hebrew University of Jerusalem *Summer 2022*

Thorsten Eisenhofer	**Visiting PhD student**
Currently Postdoctoral Fellow at TU Berlin	*Summer 2022*
Yannis Cattan	**Research Intern**
Currently Masters student at ENS Paris-Saclay (MVA)	*Summer 2022*
Roei Schuster	**Postdoctoral Fellow**
Currently CTO at Context AI	*2021-2022*
Ilia Shumailov (co-advised with Kassem Fawaz)	**Postdoctoral Fellow**
Currently Research Scientist at Google DeepMind and Junior Research Fellow at Oxford	*started Fall 2021*
Franziska Boenisch	**Research Intern**
Currently Postdoctoral Fellow at Vector Institute	*Summer 2021 - Spring 2022*
Hongyu (Charlie) Chen	**Engineering Science student**
Currently Machine Learning Engineer at Cohere.ai	*Fall 2021 - Summer 2022*
Aisha Alaagib	**Research Intern**
Currently Research Intern at MILA	*Summer 2021*
Armin Ale	**Engineering Science student**
Currently Software Engineer at Intel	*Summer 2021 - Summer 2022*
Emmy Fang (co-advised with Bo Wang) [DeepMind Scholar]	**MS student**
Currently PhD student at the University of Toronto	*Fall 2021 - Summer 2023*
Ali Shahin Shamsabadi	**Research Intern**
Currently Privacy Researcher at Brave	*Winter 2021 - Fall 2021*
Anvith Thudi	**Mathematics Specialist Undergraduate student**
Currently PhD student at the University of Toronto	*Fall 2020 - Summer 2022*
Adam Dziedzic	**Postdoctoral Fellow**
Currently Assistant Professor at CISPA Helmholtz Center for Information Security	*Fall 2020 - Summer 2023*
Natalie Dullerud	**MS student**
Currently PhD Student at Stanford	*Fall 2020 - Summer 2022*
Steven Xia (co-advised with Shurui Zhou)	**Undergraduate student**
Currently PhD student at UIUC	*Fall 2020 - Summer 2021*
Jin Zhou	**Engineering Science student**
Currently PhD student at Cornell	*Fall 2020 - Summer 2021*
Lucy Lu	**Engineering Science student**
Currently MS student at Stanford	*Fall 2020 - Summer 2021*
Marko Huang	**Engineering Science student**
Currently MS student at University of Toronto	*Fall 2020 - Summer 2021*
Gabriel Deza	**Engineering Science student**
Currently MS student at UC Berkeley	*Fall 2020 - Summer 2021*
Tejumade Afonja	**Research Intern**
Currently PhD student at Saarland University	*Summer 2020*
Ilia Shumailov	**Visiting PhD student**
Currently PhD student at University of Cambridge	*Summer 2020*
Milad Nasr (co-hosted with Nicholas Carlini)	**Google Brain Intern**
Currently Research Scientist at Google Brain	*Summer 2020*

Gabriel Deza	**Research Intern**
Currently Engineering Science student at UofT	*Summer 2020*
Lorna Licollari	**Research Intern**
Currently Engineering Science student at University of Toronto	*Summer 2020*
Pratyush Maini	**Research Intern**
Currently PhD student at CMU	*Summer 2020*
Yunxiang Zhang	**Research Intern**
Currently PhD student at Chinese University of Hong Kong	*Spring 2020*
Saina Asani	**Research Assistant**
Currently AI Researcher at Huawei	*Winter 2020 - Summer 2020*
Laura Zhukas	**Undergraduate Student Researcher**
Currently BASc student at the University of Waterloo	*Fall 2019*
Christopher Choquette-Choo	**Engineering Science student**
Currently Research Engineer at Google Brain	*Fall 2019 - Summer 2020*
Nick Jia	**Engineering Science student**
Currently PhD student at University of Toronto	*Fall 2019 - Summer 2020*
Baiwu Zhang	**MEng student**
Currently ML Engineer at Twitter	*Fall 2019 - Summer 2020*
Varun Chandrasekaran	**Visiting PhD student**
Currently Assistant Professor at UIUC	*Fall 2019*
Vinith Suriyakumar (co-advised with M. Ghassemi and A. Goldenberg)	**MS student**
Currently PhD student at MIT	*Fall 2019 - Summer 2021*
Lucas Bourtoule	**MASc student**
Currently Security Engineer at Trail of Bits	*started Fall 2019*
Adelin Travers (co-advised with David Lie)	**PhD student**
Currently Senior ML Assurance Engineer at Trail of Bits	*Fall 2019 - Summer 2021*
Hadi Abdullah (co-hosted with Damien Octeau)	**Google Intern**
Currently Researcher at Visa Research	*Summer 2019*
Matthew Jagielski	**Google Brain intern**
Currently Research Scientist at Google Brain	*Summer 2019*

Selected Professional Activities

Chair (Conferences)

SaTML: IEEE Conference on Secure and Trustworthy Machine Learning	*2023, 2024*

Associate Chair or Area Chair (Conferences)

Oakland: IEEE Symposium on Security and Privacy	*2022, 2023*
NeurIPS: Neural Information Processing Systems	*2021, 2022*

Program Committee Member (Conferences)

Oakland: IEEE Symposium on Security and Privacy	*2020, 2021, 2024*
USENIX Security: USENIX Security Symposium	*2019, 2020, 2021*
CCS: ACM Conference on Computer and Communications Security	*2018, 2019, 2020*
NeurIPS: Workshop Committee Member	*2020*

PETS: Privacy Enhancing Technologies Symposium	*2019*
NDSS: Network and Distributed System Security Symposium	*2018*

Reviewer (Conferences)

ICLR: International Conference on Learning Representations	*2019, 2020, 2021*
NeurIPS: Neural Information Processing Systems	*2017, 2018, 2020*
CHIL: ACM Conference on Health, Conference, and Learning	*2020*
ICML: International Conference on Machine Learning	*2017, 2018, 2019*
AAAI: AAAI Conference on Artificial Intelligence	*2019*
USENIX Security: USENIX Security Symposium	*2018*
Oakland: IEEE Symposium on Security and Privacy	*2017, 2018*

Action Editor (Journals)

TMLR: Transactions on Machine Learning Research	*2022*

Reviewer (Journals)

Nature	*2020*
Journal of Computer Security	*2018*
IEEE Pervasive special issue on "Securing the IoT"	*2017*
IEEE Transactions on Information Forensics and Security	*2017*
IEEE Transactions on Dependable and Secure Computing	*2017*
IEEE Security and Privacy Magazine	*2017*

Chair (Workshops)

Royal Society and Royal Society of Canada: Frontiers of Science Joint Meeting on AI	*2024*
ICLR workshop on “Towards Trustworthy ML: Rethinking Security and Privacy for ML”	*2020*
NeurIPS workshop on Security in ML	*2018*

Organizing Committee (Workshops)

Oakland (IEEE S&P) Workshop: Deep Learning and Security (DLS)	*2021*
DSN Workshop: Dependable and Secure ML	*2019-2023*
ICML Workshop: Security and Privacy of ML	*2019*
NeurIPS Competition: Adversarial ML	*2018*
NeurIPS Workshop: Secure ML	*2017*

Reviewer (Funding)

AI Xprize	*2017-2020*
Google Faculty Research Awards	*2017, 2018, 2019*
Agence Nationale de la Recherche	*2017*

Invited Participant and Consultations

Bellairs Workshop on Contemporary, Foreseeable and Catastrophic Risks of LLMs: Participant	*2024*
Witness: House of Commons of Canada	*2024*
Consultation: UK Secretary of State for Science, Innovation and Technology	*2024*
Consultation: French National Data Protection Authority (CNIL)	*2024*
Joint Assembly Canada-France Committee for Science, Technology, and Innovation: Participant	*2023*
Advisory Board for the TESTABLE EU H2020 Consortium: Member	*2022-2025*

CIFAR Pan-Canadian AI Strategy Responsible AI Working Group: Chair *2023*
CIFAR Pan-Canadian AI Strategy National Program Committee: Member *2022-2023*
Audition: Rhone-Alpes Conseil economique, social et environnemental *2022*
Briefing: Microsoft Azure CTO *2022*
Interview: French National Data Protection Authority (CNIL) *2022*
Consultation: Robert O. Work and Michele Flournoy *2021*
Consultation: National Security and Intelligence Review Agency *2021*
AI Governance Workshop: Rockefeller and Mozilla Foundations *2021*
Consultation: Nathaniel Erskine-Smith (Member of Parliament) *2021*
Consultation: Chief Privacy Officer of Ontario *2021*
Privacy and ML interest group: Alan Turing Institute *2021*
Robust Artificial Intelligence: Lorentz Center *2021*
Advisory Board Member: mytrace.ca *2020*
Privacy and ML: socml.org *2020*
Security of Machine Learning: Dagstuhl Seminar (declined due to COVID) *2020*
Consultation: Privacy Commissionner of Canada *2020*
NSTC Workshop on AI and Cybersecurity: University of Maryland *2019*
Briefing: JASON advisory group *2018*
"When Humans Attack" workshop: Data and Society Research Institute *2018*
ARO/IARPA Workshop on Adversarial Machine Learning: University of Maryland *2018*
ARO Workshop on Adversarial Machine Learning: Stanford *2017*
DARPA Workshop on Safe Machine Learning: Simons Institute *2017*

Service at the University of Toronto

Deep Learning Faculty Hiring Committee: Member *2019-2023*

Service at the Vector Institute

Faculty Hiring Committee: Chair *2023-2024*
Faculty Affiliate Hiring Committee: Chair *2022-2023*
Faculty Hiring Committee: Member *2020-2022*
Faculty Affiliate Hiring Committee: Member *2019-2022*

Keynotes, Panels and Invited Talks

Keynotes

Training Dynamics and Trust in ML: Deep Learning Indaba Trustworthy AI Workshop *2023*
What does it mean for ML to be trustworthy?: MITRE *2023*
The Role of Randomization in Trustworthy ML: ACM CCS Workshop on MTD *2022*
Is Differential Privacy a Silver Bullet for ML?: 35th Canadian Conference on AI *2022*
What does it mean for ML to be trustworthy?: CAMLIS 2021 *2021*
What does it mean for ML to be trustworthy?: ESORICS *2021*
What does it mean for ML to be trustworthy?: EVOKE CASCON 2020 *2020*
What does it mean for ML to be trustworthy?: AsiaCCS Workshop of Security and Privacy in AI *2020*

What does it mean for ML to be trustworthy?: RAISA3 at the European Conference on AI *2020*

What does it mean for ML to be trustworthy?: Samsung Security Tech Forum *2020*

What does it mean for ML to be trustworthy?: NSERC COHESA Annual Meeting *2020*

What does it mean for ML to be trustworthy?: ICML Workshop on Participatory ML *2020*

How Relevant Is the Turing Test in the Age of Sophisbots?: CVPR Workshop on Media Forensics *2020*

Security and Privacy in Machine Learning: France is AI 2019 *2019*

A Marauder's Map of Security and Privacy in ML: CVPR workshop on Privacy and Security *2019*

A Marauder's Map of Security and Privacy in ML: AISec '18 *2018*

Tutorials

Security and Privacy in ML: INRIA Data Institute *2018*

Security and Privacy in ML: IEEE WIFS 2017 *2017*

Adversarial ML with CleverHans: ODSC West (joint with Nicholas Carlini) *2017*

Adversarial ML with CleverHans: ICML workshop on Reproducibility in ML *2017*

Guest Lectures

Course on Trustworthy Machine Learning: National Yang Ming Chiao Tung University *2023*

Artificial Intelligence: Lycee Francais de Toronto *2023*

Security in Machine Learning: University of Michigan *2023*

The Role of Randomization in Trustworthy ML: University of Seoul *2023*

What does it mean for ML to be trustworthy?: University of Pittsburgh *2022*

What does it mean for ML to be trustworthy?: Korea Institute of Inf. Security and Cryptology *2022*

What does it mean for ML to be trustworthy?: Purdue University *2020*

What does it mean for ML to be trustworthy?: University of Wisconsin-Madison *2020*

Security and Privacy in Machine Learning: Los Alamos National Laboratory *2020*

The Limitations of Deep Learning in Adversarial Settings: Carnegie Mellon University *2020*

Machine Learning Security: Adversarial Examples: Stanford *2019*

A Marauder's Map of Security and Privacy in ML: UC Berkeley - CS294-131 *2019*

Security and Privacy in ML: Penn State University - CSE 543 *2017*

Invited Talks

TBD: ICLR 2024 Workshop on Reliable and Responsible Foundation Models *2024*

TBD: US National Academies of Sciences, Engineering, and Medicine *2024*

Characterizing Machine Unlearning: Vanderbilt University *2024*

Characterizing Machine Unlearning: University of Oxford *2024*

TBD: Apple *2024*

Characterizing Machine Unlearning: University of Cambridge *2024*

Towards Defining Machine Unlearning: Google DeepMind *2023*

Privacy and provenance of training data in LLMs: Inria *2023*

Training Dynamics and Trust in ML: Amazon *2023*

Training Dynamics and Trust in Machine Learning: Carnegie Mellon University *2023*

Security and Privacy in Machine Learning: CMU Bosch Institute *2023*

Privacy in LLMs: ElementAI *2023*

Trustworthy Machine Learning: Fujitsu *2023*

Challenges in Machine Unlearning: Northeastern University *2023*

Trustworthy Machine Learning: Defence Research and Development Canada *2023*

Training Dynamics and Trust in Machine Learning: Inria *2023*

Training Dynamics and Trust in Machine Learning: Anthropic *2023*

Training Dynamics and Trust in Machine Learning: UC Berkeley *2023*

What does it mean for ML to be trustworthy?: Google *2022*

Is Differential Privacy a Silver Bullet for Machine Learning?: ACM CIKM Workshop on FL *2022*

What does it mean for ML to be trustworthy?: CISPA *2022*

What does it mean to unlearn?: Georgetown University *2022*

What does it mean to unlearn?: ICML 2022 Workshop on Updatable ML *2022*

Is Differential Privacy a Silver Bullet for Machine Learning?: Alan Turing Institute *2022*

When the Curious Abandon Honesty: Federated Learning Is Not Private: Apple *2022*

Is Differential Privacy a Silver Bullet for Machine Learning?: UK Security and Privacy Seminar *2022*

Is Differential Privacy a Silver Bullet for Machine Learning?: Princeton University *2022*

Is Differential Privacy a Silver Bullet for Machine Learning?: UofT AI Conference *2022*

Is Differential Privacy a Silver Bullet for Machine Learning?: Microsoft Research Summit *2021*

What can trustworthy ML learn from cryptography?: CRYPTO Workshop on PPML *2021*

What does it mean for ML to be trustworthy?: IBM *2021*

What does it mean for ML to be trustworthy?: ICML 2021 Workshop on Socially Responsible ML *2021*

What does it mean for ML to be trustworthy?: CVPR 2021 Workshop on Adversarial ML *2021*

Three Flavors of Private Machine Learning: Google *2021*

What does it mean for ML to be trustworthy?: Turing Institute *2021*

Three Flavors of Private Machine Learning: Second AAAI Workshop of Privacy Preserving AI *2021*

What does it mean for ML to be trustworthy?: MIT *2021*

What does it mean for ML to be trustworthy?: Schwartz Reisman Institute *2021*

What does it mean for ML to be trustworthy?: University of Waterloo *2021*

What does it mean for ML to be trustworthy?: Vector Institute AI Masters Summit *2020*

What does it mean for ML to be trustworthy?: OpenMined Privacy Conference *2020*

Tempered Sigmoids for Deep Learning with Differential Privacy: Apple *2020*

PhD Career Paths (Academic v. Non-academic): Google PhD Intern Research Conference *2020*

What does it mean for ML to be trustworthy?: Vector Institute Endless Summer School *2020*

Machine Unlearning: Facebook *2020*

What does it mean for ML to be trustworthy?: USENIX Enigma *2020*

Security and Privacy in Machine Learning: King's College London *2020*

TensorFlow Privacy: TensorFlow Roadshow Paris *2019*

Security and Privacy in Machine Learning: Columbia University *2019*

Security and Privacy in Machine Learning: Fields Institute *2019*

A Marauder's Map of Security and Privacy in ML: Cybersecurity AI Prague *2019*

Security and Privacy in ML: Carleton University *2019*

A Marauder's Map of Security and Privacy in ML: Princeton University *2019*

A Marauder's Map of Security and Privacy in ML: University of British Columbia *2019*

A Marauder's Map of Security and Privacy in ML: IBM AI week security symposium *2019*

Security and Privacy in Machine Learning: Waterloo ML + Security + Verification Workshop *2019*

Machine Learning at Scale with Differential Privacy in TensorFlow: USENIX PEPR 2019 *2019*

PhD Career Paths (Academic v. Non-academic): Google PhD Intern Research Conference *2019*

PhD Career Paths (Academic v. Non-academic): Google PhD Fellowship Summit *2019*

Security and Privacy in ML: Microsoft *2019*

Security and Privacy in ML: National Academies Workshop on AI and ML for Cybersecurity *2019*

A Marauder's Map of Security and Privacy in ML: Palo Alto Networks *2019*

A Marauder's Map of Security and Privacy in ML: Google Brain Zurich *2019*

A Marauder's Map of Security and Privacy in ML: EPFL Applied ML Days *2019*

Security and Privacy in ML: Google Launchpad Studio *2018*

Security and Privacy in ML: MSR Cambridge AI Summer School *2018*

Characterizing the Space of Adversarial Examples in ML: NVIDIA *2018*

Characterizing the Space of Adversarial Examples in ML: 2nd ARO/IARPA Workshop on AML *2018*

Characterizing the Space of Adversarial Examples in ML: MIT-IBM Watson AI Lab *2018*

Characterizing the Space of Adversarial Examples in ML: MSR Cambridge *2018*

Characterizing the Space of Adversarial Examples in ML: University of Toronto *2018*

Characterizing the Space of Adversarial Examples in ML: EPFL *2018*

Characterizing the Space of Adversarial Examples in ML: University of Southern California *2018*

Characterizing the Space of Adversarial Examples in ML: University of Michigan *2018*

Characterizing the Space of Adversarial Examples in ML: MPI for Software Systems *2018*

Characterizing the Space of Adversarial Examples in ML: Columbia University *2018*

Characterizing the Space of Adversarial Examples in ML: University of Virginia *2018*

Characterizing the Space of Adversarial Examples in ML: Intel Labs *2018*

Characterizing the Space of Adversarial Examples in ML: McGill University *2018*

Characterizing the Space of Adversarial Examples in ML: University of Florida *2018*

Security and Privacy in ML: Age of AI Conference *2018*

Security and Privacy in ML: Bar Ilan University *2018*

Security and Privacy in ML: IVADO *2018*

Security and Privacy in ML: Ecole Polytechnique Montreal *2018*

Security and Privacy in ML: Element AI *2018*

Security and Privacy in ML: Georgian Partners *2017*

Private Machine Learning with PATE: With the Best online conference *2017*

Gradient Masking in ML: Stanford - ARO Adversarial ML Workshop *2017*

Security and Privacy in ML: Ecole Centrale de Lyon *2017*

Security and Privacy in ML: Oxford University *2017*

Adversarial Examples in ML: AI with the Best (joint with Patrick McDaniel) *2017*

Security and Privacy in ML: Deep Learning Summit Singapore *2017*

Security and Privacy in ML: MSR Cambridge *2017*

Security and Privacy in ML: University of Cambridge *2017*

Private Aggregation of Teacher Ensembles: Stanford *2017*

Adversarial ML: Data Mining for Cyber Security meetup *2017*
Private Aggregation of Teacher Ensembles: Symantec *2017*
Adversarial Examples in ML: Usenix Enigma 2017 *2017*
Private Aggregation of Teacher Ensembles: LeapYear *2017*
Private Aggregation of Teacher Ensembles: Immuta *2017*
Security and Privacy in ML: Ecole Centrale de Lyon *2016*
Adversarial Examples in ML: LinkedIn *2016*
Adversarial Examples in ML: Stanford *2016*
Adversarial Examples in ML: Berkeley *2016*
Adversarial Examples in ML: AutoSens (joint with Ian Goodfellow) *2016*
Adversarial Examples in ML: Google *2016*

Panels

TBD: ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models *2024*
Privacy in the world of AI: University of Toronto *2024*
Hardware for Machine Learning: MICRO2023 *2023*
EDI Panelist at Prospective Professors in Training Program: University of Toronto *2023*
Do we want to limit AI?: Schwartz Reisman Institute for Technology and Society *2023*
What does it mean to unlearn?: University of Waterloo *2022*
Robust and Reliable ML: ICLR 2021 Workshop on Robust and Reliable ML *2021*
Adversarial Examples in ML: Stanford AI Salon (joint with Ian Goodfellow) *2017*
Machine Learning and Security: NSF 2017 SaTC PIs Meeting *2017*
What role will AI play in the future of autonomous vehicles and ADAS?: AutoSens *2016*

Teaching and Community Outreach

Teaching at the University of Toronto

ECE421H: Introduction to Machine Learning *Fall 2023*
ECE1784H/CSC2559H: Trustworthy Machine Learning *Fall 2022*
ECE421H: Introduction to Machine Learning *Fall 2022*
ECE1784H/CSC2559H: Trustworthy Machine Learning *Fall 2021*
ECE421H: Introduction to Machine Learning *Fall 2021*
ECE421H: Introduction to Machine Learning *Fall 2020*
ECE1513H: Introduction to Machine Learning *Fall 2020*
ECE1513H: Introduction to Machine Learning *Winter 2020*
ECE1784H: Trustworthy Machine Learning *Fall 2019*

Software

TensorFlow Privacy: Co-author of open-source library for differentially private ML *2019*
CleverHans: Co-author of open-source library for adversarial ML *2016*

CleverHans Blog

We need a 21st century framework for 21st century problems *2022*
Can stochastic pre-processing defenses protect your models? *2022*

Are adversarial examples against proof-of-learning adversarial?	*2022*
How to Keep a Model Stealing Adversary Busy?	*2022*
All You Need Is Matplotlib	*2022*
How to deploy machine learning with differential privacy? (DifferentialPrivacy.org)	*2021*
Arbitrating the integrity of stochastic gradient descent with proof-of-learning	*2021*
Beyond federation: collaborating in ML with confidentiality and privacy	*2021*
Is this model mine?	*2021*
Why we should regulate information about persons, not personal information	*2021*
To guarantee privacy, focus on the algorithms, not the data	*2021*
Teaching Machines to Unlearn	*2020*
In Model Extraction, Don't Just Ask How?: Ask Why?	*2020*
How to steal modern NLP systems with gibberish?	*2020*
The academic job search for computer scientists in 10 questions	*2019*
How to know when machine learning does not know	*2019*
Machine Learning with Differential Privacy in TensorFlow	*2019*
Privacy and machine learning: two unexpected allies?	*2018*
The challenge of verification and testing of machine learning	*2017*
Is attacking machine learning easier than defending it?	*2017*
Breaking things is easy	*2016*

Selected Media Coverage

The Economist. AI could accelerate scientific fraud as well as progress

The Register. What is Model Collapse and how to avoid it

Prospect Magazine. What happens when AI trains itself?

Wired. Confessions of a Viral AI Writer

Le Monde. L intelligence artificielle peut-elle s effondrer sur elle-meme ?

UofT News. Training AI on machine-generated text could lead to 'model collapse,' researchers warn

Financial Times. The sceptical case on generative AI

Independent. Scientists warn of threat to internet from AI-trained AIs

IEEE Spectrum. The Internet Isn't Completely Weird Yet; AI Can Fix That

Financial Times. Why computer-made data is being used to train AI models

New York Times. Wikipedia's Moment of Truth

Wall Street Journal. AI Junk Is Starting to Pollute the Internet

New York Times Podcast - Hard Fork. Is A.I. Poisoning Itself?

Schneier on Security. Class-Action Lawsuit for Scraping Data without Permission

The Atlantic. AI Is an Existential Threat to Itself

Business Insider. AIs trained on other AI output will start producing junk within a few generations, scientists warn

Cosmos Magazine. Degenerative AI: Researchers say training artificial intelligence models on machine-generated data leads to model collapse

NewScientist. AIs will become useless if they keep learning from other AIs

VentureBeat. The AI feedback loop: Researchers warn of model collapse as AI trains on AI-generated content

BBC. Can Artificial Intelligence teach itself?

Global News. Ontario urged to develop guardrails on public sector use of AI

TFO. L intelligence artificielle va-t-elle redefinir l apprentissage ?

RadioCanada. Les robots conversationnels tels que ChatGPT

CBC. As new AI ChatGPT earns hype, cybersecurity experts warn about potential malicious uses

RadioCanada. Interview for Chronique CERVO FRANCO

Schneier on Security. Attacking the Performance of Machine Learning Systems

Schneier on Security. Manipulating Machine-Learning Systems through the Order of the Training Data

CACM. Can AI Learn to Forget?

New York Times. As Hackers Take Down Newfoundland's Health Care System, Silence Descends

RadioCanada. Ottawa finance la creation d'un outil pour dechiffrer les mots de passe

Wired. Now That Machines Can Learn, Can They Unlearn?

The Register. Hey, AI software developers, you are taking Unicode into account, right ... right?

TechSequences Podcast. Can advances in technology help liberate us from the grip of disinformation?

Heise.de. Machine Unlearning: Algorithmen können nichts vergessen

VentureBeat. How adversarial attacks reveal machine learning's weakness

DeepLearning.Ai. about adv x contest

Quartz. OpenAI has a new tool that could keep hackers from wrecking a self-driving car

Quartz. AI can learn from data without ever having access to it

Communications of the ACM. Learning Securely

Wired. How to Steal an AI

Popular Science. Fooling the machine

Die Zeit. Notwehr against the machine

Fast Company. How To Fool A Neural Network

TheVerge. Magic AI: these are the optical illusions that trick, fool, and flummox computers

GCN. Machines learning evolves, and hackers stand to gain

MIT Technology Review. Human brains and AIs can be hacked with these weird tweaked photos

TheNextWeb. Google teaches AI to fool humans so it can learn from our mistakes

IEEE Spectrum. Hacking the Brain With Adversarial Images

TWiML. Scalable Differential Privacy for Deep Learning with Nicolas Papernot

Le Monde. Les bugs de l'intelligence artificielle

The Verge. Google is making it easier for AI developers to keep users' data private

VentureBeat. Google introduces TensorFlow Privacy, a machine learning library with strong privacy guarantees