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**
Assistant Professor in the Department of Electrical & Computer Engineering — *Since 09/2019*

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

Google Brain — **Mountain View, CA**
Research Scientist — *Since 08/2018*

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

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

Education

Pennsylvania State University — **University Park, PA**
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**
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) — *2010–2012*

Honors

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*
Research Assistantship: Pennsylvania State University *2014–2015*
Scholarship for Exceptional Academic Achievements: McGill *2010*

Publications

Pre-prints

High-Fidelity Extraction of Neural Network Models. *Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, Nicolas Papernot*. (2019)

Prototypical Examples in Deep Learning: Metrics, Characteristics, and Utility. *Nicholas Carlini, Ulfar Erlingsson, Nicolas Papernot*. (2019)

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

Conference proceedings

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]** (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 Networksg. *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)

Workshop publications

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. *H. 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. *P. 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

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

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)

Invited publications

How Relevant Is the Turing Test in the Age of Sophisbots?. *Dan Boneh, Andrew J. Grotto, Patrick McDaniel, Nicolas Papernot*. To appear in 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)

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)

Professional Activities

Chair

NeurIPS workshop on Security in ML *2018*

Organizing Committee

ICML Workshop on the Security and Privacy of ML *2019*

DSN Workshop on Dependable and Secure ML *2019*

NeurIPS Competition on Adversarial ML *2018*

NeurIPS Workshop on Secure ML *2017*

Program Committee (Conferences)

Oakland: IEEE Symposium on Security and Privacy *2020*

USENIX Security: USENIX Security Symposium *2019, 2020*

CCS: ACM Conference on Computer and Communications Security *2018, 2019*

PETS: Privacy Enhancing Technologies Symposium *2019*

ACSAC: Annual Computer Security Applications Conference *2018*

AsiaCCS: ACM Asia Conference on Computer and Communications Security *2018*

GameSec: Conference on Decision and Game Theory for Security *2018*

NDSS: Network and Distributed System Security Symposium *2018*

Program Committee (Workshops)

DLS: Deep Learning and Security colocated with Oakland *2018*

CV-COPS: Privacy and Security colocated with CVPR *2018*

DSML: Dependable and Secure ML colocated with DSN *2018*

Reviewer (Conferences)

ICML: International Conference on Machine Learning *2017, 2018, 2019*
ICLR: International Conference on Learning Representations *2019, 2020*
AAAI: AAAI Conference on Artificial Intelligence *2019*
USENIX Security: USENIX Security Symposium *2018*
Oakland: IEEE Symposium on Security and Privacy *2017, 2018*
NeurIPS: Neural Information Processing Systems *2017, 2018*
ACM WiSec: ACM Conference on Security and Privacy in Wireless and Mobile Networks *2016*
DIMVA: Conference on Detection of Intrusions and Malware and Vulnerability Assessment *2016*

Reviewer (Journals)

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*

Reviewer (Funding)

Agence Nationale de la Recherche *2017*
AI Xprize *Since 2017*
Google Faculty Research Awards *2017, 2018*

Invited Participant

NSTC Workshop on AI and Cybersecurity: University of Maryland *2019*
"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*

Defense Committee

Ryan Sheatsley: MSc, Pennsylvania State University *2018*

Keynotes, Panels and Invited Talks

Keynotes

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

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

TensorFlow Privacy: TensorFlow Roadshow Paris *2019*
Title TBD: 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*
Title TBD: France is AI 2019 *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*
Title TBD: 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: *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

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*

Students

Varun Chandrasekaran: visiting PhD, Fall 2019
Lucas Bourtoule: MASc, starting Fall 2019
Adelin Travers: PhD, starting Fall 2019, co-supervised with David Lie
Matthew Jagielski: Google Brain intern, Summer 2019

Teaching and Community Outreach

Teaching at the University of Toronto

ECE1784H: Trustworthy Machine Learning *Fall 2019*

Teaching at Pennsylvania State University

CSE 597: Advanced Topics in the Security and Privacy of ML: Co-instructor *2017*
CSE 597: Security and Privacy of Machine Learning: Co-instructor *2016*

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 (co-authored with Ian Goodfellow)

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*