Florian Tramèr

ETH Zürich
Department of Computer Science
8006 Zürich, Switzerland
✉ florian.tramer@inf.ethz.ch
🏠 http://www.floriantramer.com

Current Positions

ETH Zürich, Zürich, Switzerland 2022 - present
Assistant Professor of Computer Science

Invariant Labs, Zürich, Switzerland 2024 - present
Co-founder & Academic advisor

Education

Stanford University, Stanford CA, USA 2016 - 2021
Ph.D. in Computer Science
Advisor: Prof. Dan Boneh
Thesis: “Measuring and Enhancing the Security of Machine Learning’

EPFL, Lausanne, Switzerland 2013 - 2015
M.Sc in Computer Science
Advisor: Prof. Jean-Pierre Hubaux
Thesis: “Algorithmic Fairness Revisited”

EPFL, Lausanne, Switzerland 2009 - 2012
B.Sc in Computer Science
Exchange year (2011-2012), Carnegie Mellon University, Pittsburgh PA, USA

Research Appointments

Google Brain, Zürich, Switzerland 2021 - 2022
Visiting Faculty (hosted by Dr. Andreas Terzis)

Chainlink, San Francisco View, CA, USA 2020 - 2021
Research Consultant (advised by Prof. Ari Juels)

Google, Mountain View, CA, USA June - Sep. 2018
Software Engineering Intern (advised by Dr. Sai Deep Tetali)

IBM Research, Yorktown Heights, NY, USA June - Sep. 2017
Research Intern (advised by Dr. Evelyn Duesterwald)

EPFL, Lausanne, Switzerland 2015 - 2016
Scientific Assistant (advised by Prof. Jean-Pierre Hubaux)

EPFL, Lausanne, Switzerland 2013 - 2015
Research Assistant (advised by Prof. Serge Vaudenay)

Teaching and Advising

Classes.

Privacy Enhancing Technologies, ETHZ, 2024-
Applied Cryptography, ETHZ, 2024- (co-Instructor)
Large Language Models, ETHZ, 2023- (co-Instructor)
Information Security Lab, ETHZ, 2022- (co-Instructor)

CS355: Topics in Cryptography, Stanford, 2018 (Teaching Assistant), 2019 & 2020 (Instructor)

Current Group.

Avital Shafran (Postdoc)
Edoardo Debenedetti
Daniel Paleka
Javier Rando → Anthropic
Michael Aerni
Jie Zhang
Kristina Nikolic
Lukas Fluri
Pura Peetathawatchai

Invited Lectures.

University of Luxembourg, *Poisoning and Misusing ML Models*, AI & Security, 2023
University of Michigan, *Security and Privacy in Machine Learning*, EECS388 Intro to Security, 2022
Stanford, *Breaking and Safeguarding Privacy in Machine Learning*, CS356 Topics in Computer and Network Security, 2022
ETHZ, *Security and Privacy in Machine Learning* , Information Security Lab, 2021
Stanford, *Don't use Computer Vision for Web Security*, CS356 Topics in Computer and Network Security, 2020
Stanford, *Integrity and Confidentiality in Machine Learning*, CS521 Seminar on AI Safety, 2018
Stanford, *Security for Smart Contracts*, CS359B Designing Decentralized Applications, 2018
EPFL, *Fairness in Algorithmic Decision Making*, Privacy Protection, 2016

Selected Awards

Winner of SafeBench competition, 2025.

Best Paper Award, NeurIPS SoLaR Workshop 2024.

Best Paper Award, ICML 2024.

Best Paper Award, ICML 2024.

Best Paper Award runner-up, SaTML 2024.

Amazon Research Award, 2024

Google Research Scholar Award, 2024

Best Paper Finalist, INLG 2023.

Distinguished Paper Award, USENIX Security 2023.

Caspar Bowden Award runner-up, 2023.

Best Paper Award, NeurIPS ML Safety Workshop 2022.

AdvML Rising Star Award, 2021.

Best Paper Award, ICML AdvML Workshop 2021.

EPFL Master Award (Highest EPFL GPA for complete Master studies), 2015.

Scholarships and Grants

Schmidt Sciences, 2024
FAR AI, 2024
AI Safety Fund, 2024
Open Philanthropy Project, 2024
Lightspeed, 2024
Foresight AI Safety Grant, 2024
SNSF Project Grant, 2023
Open Philanthropy Project Gift, 2018
SNSF DOC.Mobility Fellowship, 2018 - 2020
Bakala Foundation Fellowship, 2016
EPFL Excellence Fellowship (Awarded for outstanding academic records), 2013 - 2015
EPFL Research Scholars MSc Program, 2013 - 2015

Publications

Journal Articles

[J4] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaid Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Mariana Raykova, Hang Qi, Daniel Ramage, Ramesh Raskar, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. "Advances and Open Problems in Federated Learning". *Foundations and Trends in Machine Learning* (Jan. 2021).

[J3] Lorenz Breidenbach*, Phil Daian*, Florian Tramèr*, and Ari Juels. "The Hydra Framework for Principled, Automated Bug Bounties". *IEEE Security & Privacy* 17.4 (July 2019), pp. 53–61. (*joint first authors).

[J2] Jean Louis Raisaro*, Florian Tramèr*, Zhanglong Ji*, Diyue Bu*, Yongan Zhao, Knox Carey, et al. "Addressing Beacon Re-Identification Attacks: Quantification and Mitigation of Privacy Risks". *Journal of the American Medical Informatics Association (JAMIA)* 24.4 (Feb. 2017), pp. 799–805. (*joint first authors).

[J1] Sonia Bogos, Florian Tramèr, and Serge Vaudenay. "On Solving LPN using BKW and Variants". *Cryptography and Communications* 8.3 (July 2016), pp. 331–369. (alphabetical author ordering).

Conference Proceedings

[C68] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. "Defeating Prompt Injections by Design". In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Apr. 2026.

[C67] Jie Zhang, Cezara Petrui, Kristina Nikolić, and Florian Tramèr. "RealMath: A Continuous Benchmark for Evaluating Language Models on Research-Level Mathematics". In *Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2025.

[C66] Nicholas Carlini, Javier Rando, Edoardo Debenedetti, Milad Nasr, and Florian Tramèr. "AutoAdvExBench: Benchmarking autonomous exploitation of adversarial example defenses". In *International Conference on Machine Learning (ICML)*. Oral Presentation. July 2025.

[C65] Yangsibo Huang, Milad Nasr, Anastasios Angelopoulos, Nicholas Carlini, Wei-Lin Chiang, Christopher A. Choquette-Choo, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Ken Ziyu Liu, Ion Stoica, Florian Tramèr, and Chiyuan Zhang. "Exploring and mitigating adversarial manipulation of voting-based leaderboards". In *International Conference on Machine Learning (ICML)*. Oral Presentation. July 2025.

[C64] Kristina Nikolić, Luze Sun, Jie Zhang, and Florian Tramèr. "The Jailbreak Tax: How Useful are Your Jailbreak Outputs?" In *International Conference on Machine Learning (ICML)*. Spotlight. July 2025.

[C63] Xuandong Zhao, Sam Gunn, Miranda Christ, Jaiden Fairoze, Andres Fabrega, Nicholas Carlini, Sanjam Garg, Sanghyun Hong, Milad Nasr, Florian Tramèr, et al. "SoK: Watermarking for AI-Generated Content". In *IEEE Symposium on Security and Privacy (S&P)*. May 2025.

[C62] Michael Aerni*, Javier Rando*, Edoardo Debenedetti, Nicholas Carlini, Daphne Ippolito, and Florian Tramèr. "Measuring Non-Adversarial Reproduction of Training Data in Large Language Models". In *International Conference on Learning Representations (ICLR)*. Apr. 2025. (*joint first authors).

[C61] Robert Hönig, Javier Rando, Nicholas Carlini, and Florian Tramèr. "Adversarial Perturbations Cannot Reliably Protect Artists From Generative AI". In *International Conference on Learning Representations (ICLR)*. Spotlight Presentation. Apr. 2025.

[C60] Milad Nasr*, Javier Rando*, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Florian Tramèr, and Katherine Lee. "Scalable Extraction of Training Data from (Production) Language Models". In *International Conference on Learning Representations (ICLR)*. Apr. 2025. (*joint first authors).

[C59] Fredrik Nestaas, Edoardo Debenedetti, and Florian Tramèr. "Adversarial Search Engine Optimization for Large Language Models". In *International Conference on Learning Representations (ICLR)*. Apr. 2025.

[C58] Daniel Paleka*, Abhimanyu Pallavi Sudhir*, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. "Consistency Checks for Language Model Forecasters". In *International Conference on Learning Representations (ICLR)*. Oral Presentation. Apr. 2025. (*joint first authors).

[C57] Yiming Zhang*, Javier Rando*, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. "Persistent Pre-Training Poisoning of LLMs". In *International Conference on Learning Representations (ICLR)*. Apr. 2025. (*joint first authors).

[C56] Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. "Membership Inference Attacks Cannot Prove that a Model Was Trained On Your Data". In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Apr. 2025.

[C55] Jakub Łucki, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr, and Javier Rando. "An adversarial perspective on machine unlearning for AI safety". In *Transaction on Machine Learning Research (TMLR)*. Mar. 2025.

[C54] Patrick Chao*, Edoardo Debenedetti*, Alexander Robey*, Maksym Andriushchenko*, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramèr, et al. "JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models". In *Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2024. (*joint first authors).

[C53] Edoardo Debenedetti*, Javier Rando*, Daniel Paleka*, Silaghi Fineas Florin, Dragos Albastroiu, Niv Cohen, Yuval Lemberg, Reshmi Ghosh, Rui Wen, Ahmed Salem, Giovanni Cherubin, Santiago Zanella-Beguelin, Robin Schmid, Victor Klemm, Takahiro Miki, Chenhao Li, Stefan Kraft, Mario Fritz, Florian Tramèr, Sahar Abdelnabi, and Lea Schönherr. "Dataset and Lessons Learned from the 2024 SaTML LLM Capture-the-Flag Competition". In *Conference on Neural Information Processing Systems (NeurIPS)*. Spotlight Presentation. Dec. 2024. (*joint first authors).

[C52] Edoardo Debenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. "AgentDojo: A Dynamic Environment to Evaluate Attacks and Defenses for LLM Agents". In *Conference on Neural Information Processing Systems (NeurIPS)*. Winner of CAIS SafeBench competition. Dec. 2024.

[C51] Jonathan Hayase, Ema Borevkovic, Nicholas Carlini, Florian Tramèr, and Milad Nasr. "Query-Based Adversarial Prompt Generation". In *Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2024.

[C50] Michael Aerni*, Jie Zhang*, and Florian Tramèr. "Evaluations of Machine Learning Privacy Defenses are Misleading". In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Oct. 2024. (*joint first authors).

[C49] Edoardo Debenedetti, Giorgio Severi, Nicholas Carlini, Christopher A Choquette-Choo, Matthew Jagielski, Milad Nasr, Eric Wallace, and Florian Tramèr. "Privacy Side Channels in Machine Learning Systems". In *USENIX Security Symposium*. Aug. 2024.

[C48] Nicholas Carlini, Daniel Paleka, Krishnamurthy Dj Dvijotham, Thomas Steinke, Jonathan Hayase, A Feder Cooper, Katherine Lee, Matthew Jagielski, Milad Nasr, Arthur Conmy, Eric Wallace, David Rolnick, and Florian Tramèr. "Stealing Part of a Production Language Model". In *International Conference on Machine Learning (ICML)*. Best Paper Award. July 2024.

[C47] Shanglun Feng and Florian Tramèr. "Privacy Backdoors: Stealing Data with Corrupted Pretrained Models". In *International Conference on Machine Learning (ICML)*. July 2024.

[C46] Francesco Pinto, Nathalie Rauschmayr, Florian Tramèr, Philip Torr, and Federico Tombari. "Extracting Training Data From Document-Based VQA Models". In *International Conference on Machine Learning (ICML)*. July 2024.

[C45] Florian Tramèr, Gautam Kamath, and Nicholas Carlini. "Position: Considerations for Differentially Private Learning with Large-Scale Public Pretraining". In *International Conference on Machine Learning (ICML)*. Best Paper Award. July 2024. (reverse-alphabetical author ordering).

[C44] Javier Rando and Florian Tramèr. "Universal Jailbreak Backdoors from Poisoned Human Feedback". In *International Conference on Learning Representations (ICLR)*. May 2024.

[C43] Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. "Poisoning Web-Scale Training Datasets is Practical". In *IEEE Symposium on Security and Privacy (S&P)*. May 2024.

[C42] Edoardo Debenedetti, Nicholas Carlini, and Florian Tramèr. "Evading Black-box Classifiers Without Breaking Eggs". In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Best paper award runner-up. Apr. 2024.

[C41] Lukas Fluri*, Daniel Paleka*, and Florian Tramèr. "Evaluating Superhuman Models with Consistency Checks". In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. Apr. 2024. (*joint first authors).

[C40] Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Anas Awadalla, Pang Wei Koh, Daphne Ippolito, Katherine Lee, Florian Tramèr, and Ludwig Schmidt. "Are aligned neural networks adversarially aligned?" In *Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2023.

[C39] Matthew Jagielski, Milad Nasr, Christopher Choquette-Choo, Katherine Lee, Nicholas Carlini, and Florian Tramèr. "Students Parrot Their Teachers: Membership Inference on Model Distillation". In *Conference on Neural Information Processing Systems (NeurIPS)*. Oral Presentation. Dec. 2023.

[C38] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. "Counterfactual Memorization in Neural Language Models". In *Conference on Neural Information Processing Systems (NeurIPS)*. Spotlight Presentation. Dec. 2023.

[C37] Daphne Ippolito, Florian Tramèr, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher A Choquette-Choo, and Nicholas Carlini. "Preventing Verbatim Memorization in Language Models Gives a False Sense of Privacy". In *International Natural Language Generation Conference (INLG)*. Best Paper Finalist. Sept. 2023.

[C36] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. "Extracting Training Data from Diffusion Models". In *USENIX Security Symposium*. Aug. 2023.

[C35] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. "Tight Auditing of Differentially Private Machine Learning". In *USENIX Security Symposium*. Distinguished paper award. Aug. 2023.

[C34] Chawin Sitawarin, Florian Tramèr, and Nicholas Carlini. "Preprocessors Matter! Realistic Decision-Based Attacks on Machine Learning Systems". In *International Conference on Machine Learning (ICML)*. July 2023.

[C33] Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramèr, and Chiyuan Zhang. "Quantifying Memorization Across Neural Language Models". In *International Conference on Learning Representations (ICLR)*. Spotlight Presentation. May 2023. (alphabetical author ordering).

[C32] Nicholas Carlini*, Florian Tramèr*, Rice Leslie, Mingjie Sun, Krishnamurthy Dvijotham, and J Zico Kolter. "(Certified!!) Adversarial Robustness for Free!" In *International Conference on Learning Representations (ICLR)*. May 2023. (*joint first authors).

[C31] Matthew Jagielski, Om Thakkar, Florian Tramèr, Daphne Ippolito, Katherine Lee, Nicholas Carlini, Eric Wallace, Shuang Song, Abhradeep Thakurta, Nicolas Papernot, and Chiyuan Zhang. "Measuring Forgetting of Memorized Training Examples". In *International Conference on Learning Representations (ICLR)*. May 2023.

[C30] Harsh Chaudhari, John Abascal, Alina Oprea, Matthew Jagielski, Florian Tramèr, and Jonathan Ullman. "SNAP: Efficient Extraction of Private Properties with Poisoning". In *IEEE Symposium on Security and Privacy (S&P)*. May 2023.

[C29] Nicholas Carlini, Matthew Jagielski, Nicolas Papernot, Andreas Terzis, Florian Tramèr, and Chiyuan Zhang. "The Privacy Onion Effect: Memorization is Relative". In *Conference on Neural Information Processing Systems (NeurIPS)*. Nov. 2022. (alphabetical author ordering).

[C28] Roland S. Zimmermann, Wieland Brendel, Florian Tramèr, and Nicholas Carlini. "Increasing Confidence in Adversarial Robustness Evaluations". In *Conference on Neural Information Processing Systems (NeurIPS)*. Nov. 2022.

[C27] Florian Tramèr, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. "Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets". In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Nov. 2022. (reverse-alphabetical author ordering).

[C26] Florian Tramèr. "Detecting Adversarial Examples Is (Nearly) As Hard As Classifying Them". In *International Conference on Machine Learning (ICML)*. Long Presentation. July 2022.

[C25] Hannah Brown, Katherine Lee, Fatemehsadat Mireshghallah, Reza Shokri, and Florian Tramèr. "What Does it Mean for a Language Model to Preserve Privacy?" In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. June 2022. (alphabetical author ordering).

[C24] Xuechen Li, Florian Tramèr, Percy Liang, and Tatsunori Hashimoto. "Large Language Models Can Be Strong Differentially Private Learners". In *International Conference on Learning Representations (ICLR)*. Oral Presentation. May 2022.

[C23] Evani Radiya-Dixit, Sanghyun Hong, Nicholas Carlini, and Florian Tramèr. "Data Poisoning Won't Save You From Facial Recognition". In *International Conference on Learning Representations (ICLR)*. May 2022.

[C22] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. "Membership Inference Attacks From First Principles". In *IEEE Symposium on Security and Privacy (S&P)*. May 2022. (alphabetical author ordering).

[C21] Mani Malek, Ilya Mironov, Karthik Prasad, Igor Shilov, and Florian Tramèr. "Antipodes of Label Differential Privacy: PATE and ALIBI". In *Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2021. (alphabetical author ordering).

[C20] Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. "Extracting Training Data from Large Language Models". In *USENIX Security Symposium*. Caspar Bowden Award for Outstanding Research in Privacy Enhancing Technologies runner-up. Aug. 2021.

[C19] Christopher A. Choquette Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. "Label-Only Membership Inference Attacks". In *International Conference on Machine Learning (ICML)*. July 2021.

[C18] Florian Tramèr and Dan Boneh. "Differentially Private Learning Needs Better Features (or Much More Data)". In *International Conference on Learning Representations (ICLR)*. Spotlight Presentation. May 2021.

[C17] Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramèr. "Is Private Learning Possible with Instance Encoding?" In *IEEE Symposium on Security and Privacy (S&P)*. May 2021. (alphabetical author ordering).

[C16] Charlie Hou, Mingxun Zhou, Yan Ji, Phil Daian, Florian Tramèr, Giulia Fanti, and Ari Juels. "SquirRL: Automating Attack Discovery on Blockchain Incentive Mechanisms with Deep Reinforcement Learning". In *Network and Distributed System Security Symposium (NDSS)*. Feb. 2021.

[C15] Florian Tramèr*, Nicholas Carlini*, Wieland Brendel*, and Aleksander Madry. "On Adaptive Attacks to Adversarial Example Defenses". In *Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2020. (*joint first authors).

[C14] Florian Tramèr, Dan Boneh, and Kenneth G. Paterson. "Remote Side-Channel Attacks on Anonymous Transactions". In *USENIX Security Symposium*. Aug. 2020, pp. 2739–2756.

[C13] Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. "Fundamental Tradeoffs between Invariance and Sensitivity to Adversarial Perturbations". In *International Conference on Machine Learning (ICML)*. July 2020.

[C12] Florian Tramèr and Dan Boneh. "Adversarial Training and Robustness for Multiple Perturbations". In *Conference on Neural Information Processing Systems (NeurIPS)*. Spotlight Presentation. Dec. 2019, pp. 5866–5876.

[C11] Florian Tramèr, Pascal Dupré, Gili Rusak, Giancarlo Pellegrino, and Dan Boneh. "AdVersarial: Perceptual Ad Blocking meets Adversarial Machine Learning". In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Nov. 2019, pp. 2005–2021.

[C10] Florian Tramèr and Dan Boneh. "Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware". In *International Conference on Learning Representations (ICLR)*. Oral Presentation. May 2019.

[C9] Lorenz Breidenbach*, Phil Daian*, Florian Tramèr*, and Ari Juels. "Enter the Hydra: Towards Principled Bug Bounties and Exploit-Resistant Smart Contracts". In *USENIX Security Symposium*. Invited to appear in IEEE Security and Privacy Magazine. Aug. 2018, pp. 1335–1352. (*joint first authors).

[C8] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. "Ensemble Adversarial Training: Attacks and Defenses". In *International Conference on Learning Representations (ICLR)*. Apr. 2018.

[C7] Anh Pham, Italo Dacosta, Bastien Jacot-Guillarmod, Kévin Huguenin, Taha Hajar, Florian Tramèr, and Jean-Pierre Hubaux. "PrivateRide: A Privacy-Preserving and Secure Ride-Hailing Service". In *Privacy Enhancing Technologies Symposium (PETS)*. July 2017, pp. 38–56.

[C6] Rafael Pass, Elaine Shi, and Florian Tramèr. "Formal Abstractions for Attested Execution Secure Processors". In *International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. Springer, Apr. 2017, pp. 260–289. (alphabetical author ordering).

[C5] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. "FairTest: Discovering Unwarranted Associations in Data-Driven Applications". In *IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, Apr. 2017, pp. 401–416.

[C4] Florian Tramèr, Fan Zhang, Huang Lin, Jean-Pierre Hubaux, Ari Juels, and Elaine Shi. "Sealed-Glass Proofs: Using Transparent Enclaves to Prove and Sell Knowledge". In *IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, Apr. 2017, pp. 19–34.

[C3] Florian Tramèr, Fan Zhang, Ari Juels, Michael Reiter, and Thomas Ristenpart. "Stealing Machine Learning Models via Prediction APIs". In *USENIX Security Symposium*. Aug. 2016, pp. 601–618.

[C2] Florian Tramèr, Zhicong Huang, Jean-Pierre Hubaux, and Erman Ayday. "Differential Privacy with Bounded Priors: Reconciling Utility and Privacy in Genome-Wide Association Studies". In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM. Oct. 2015, pp. 1286–1297.

[C1] Alexandre Duc, Florian Tramèr, and Serge Vaudenay. "Better Algorithms for LWE and LWR". In *International Conference on the Theory and Applications of Cryptographic Techniques (EUROCRYPT)*. Springer, Apr. 2015, pp. 173–202. (alphabetical author ordering).

Workshops

[W9] Debeshee Das, Jie Zhang, and Florian Tramèr. "Blind Baselines Beat Membership Inference Attacks for Foundation Models". In *Deep Learning and Security Workshop*. May 2025.

[W8] Javier Rando*, Jie Zhang*, Nicholas Carlini, and Florian Tramèr. "Adversarial ML Problems Are Getting Harder to Solve and to Evaluate". In *Deep Learning and Security Workshop*. May 2025. (*joint first authors).

[W7] Lorenzo Rossi, Michael Aerni, Jie Zhang, and Florian Tramèr. "Membership Inference Attacks on Sequence Models". In *Deep Learning and Security Workshop*. Best Paper Award. May 2025.

[W6] Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. "Backdoor Attacks for In-Context Learning with Language Models". In *ICML Workshop on Adversarial Machine Learning*. July 2023.

[W5] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramèr. "Red-Teaming the Stable Diffusion Safety Filter". In *NeurIPS Workshop on Machine Learning Safety*. Best Paper Award. Dec. 2022.

[W4] Nicholas Carlini, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Florian Tramèr. "NeuraCrypt is not private". In *CRYPTO Workshop on Privacy-Preserving Machine Learning*. Aug. 2021. (alphabetical author ordering).

[W3] Edward Chou, Florian Tramèr, and Giancarlo Pellegrino. "SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems". In *Deep Learning and Security Workshop*. May 2020.

[W2] Jörn-Henrik Jacobsen, Jens Behrmann, Nicholas Carlini, Florian Tramèr, and Nicolas Papernot. "Exploiting Excessive Invariance caused by Norm-Bounded Adversarial Robustness". In *ICLR Workshop on Safe Machine Learning*. May 2019.

[W1] Kevin Eykholt*, Ivan Evtimov*, Earlence Fernandes*, Bo Li*, Amir Rahmati*, Florian Tramèr*, Atul Prakash, Tadayoshi Kohno, and Dawn Song. "Physical Adversarial Examples for Object Detectors". In *USENIX Workshop on Offensive Technologies (WOOT)*. Aug. 2018. (*joint first authors).

Manuscripts

[M10] Milad Nasr, Nicholas Carlini, Chawin Sitawarin, Sander V. Schulhoff, Jamie Hayes, Michael Ilie, Juliette Pluto, Shuang Song, Harsh Chaudhari, Ilia Shumailov, Abhradeep Thakurta, Kai Yuanqing Xiao, Andreas Terzis, and Florian Tramèr. "The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against Llm Jailbreaks and Prompt Injections". arXiv preprint arXiv:2510.09023. Oct. 2025.

[M9] Jie Zhang, Meng Ding, Yang Liu, Jue Hong, and Florian Tramèr. "Black-box Optimization of LLM Outputs by Asking for Directions". arXiv preprint arXiv:2510.16794. Oct. 2025.

[M8] Luca Beurer-Kellner, Beat Buesser Ana-Maria Creţu, Edoardo Debenedetti, Daniel Dobos, Daniel Fabian, Marc Fischer, David Froelicher, Kathrin Grosse, Daniel Naeff, Ezinwanne Ozoani, Andrew Paverd, Florian Tramèr, and Václav Volhejn. "Design Patterns for Securing LLM Agents against Prompt Injections". arXiv preprint arXiv:2506.08837. June 2025.

[M7] Daniel Paleka, Shashwat Goel, Jonas Geiping, and Florian Tramèr. "Pitfalls in Evaluating Language Model Forecasters". arXiv preprint arXiv:2506.00723. June 2025.

[M6] Nicholas Carlini, Milad Nasr, Edoardo Debenedetti, Barry Wang, Christopher A Choquette-Choo, Daphne Ippolito, Florian Tramèr, and Matthew Jagielski. "LLMs unlock new paths to monetizing exploits". arXiv preprint arXiv:2505.11449. May 2025.

[M5] Javier Rando, Hannah Korevaar, Erik Brinkman, Ivan Evtimov, and Florian Tramèr. "Gradient-based Jailbreak Images for Multimodal Fusion Models". arXiv preprint arXiv:2410.03489. Oct. 2024.

[M4] Keane Lucas, Matthew Jagielski, Florian Tramèr, Lujo Bauer, and Nicholas Carlini. "Randomness in ML Defenses Helps Persistent Attackers and Hinders Evaluators". arXiv preprint arXiv:2302.13464. Feb. 2023.

[M3] Florian Tramèr, Andreas Terzis, Thomas Steinke, Shuang Song, Matthew Jagielski, and Nicholas Carlini. "Debugging Differential Privacy: A Case Study for Privacy Auditing". arXiv preprint arXiv:2202.12219. Feb. 2022. (reverse-alphabetical author ordering).

[M2] Rishi Bommasani et al. "On the Opportunities and Risks of Foundation Models". arXiv preprint arXiv:2108.07258. Aug. 2021.

[M1] Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. "The Space of Transferable Adversarial Examples". arXiv preprint arXiv:1704.03453. Apr. 2017.

Professional Service

Chair.
Casper Bowden PET Award, 2024
ACM CCS Workshop on Artificial Intelligence and Security (AISec), 2022 - 2024
ICLR Workshop on Trustworthy ML, 2020
NeurIPS Workshop on Security in Machine Learning, 2018

Organizing committee.
ICLR Workshop on Privacy Regulation and Protection in Machine Learning, 2024
ICML Workshop on Challenges in Deployable Generative AI, 2023
CVPR Workshop on the Art of Robustness, 2022
ICML Workshop on the Security and Privacy of Machine Learning, 2019
IEEE DSN Workshop on Dependable and Secure Machine Learning, 2019 - 2020

Area Chair.
IEEE Symposium on Security and Privacy (IEEE S&P), 2026 - present
International Conference on Machine Learning (ICML), 2025 - present
International Conference on Learning Representations (ICLR), 2024 - present
Neural Information Processing Systems (NeurIPS), 2023 - present
Transactions on Machine Learning Research (TMLR), 2023
International Conference on Artificial Intelligence and Statistics (AISTATS), 2022
Asian Conference on Machine Learning (ACML), 2022

Program committee.
IEEE Symposium on Security and Privacy (IEEE S&P), 2022 - 2025
IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), 2023 - 2024
USENIX Security Symposium, 2021 - present
ACM Conference on Computer and Communications Security (CCS), 2022
Privacy Enhancing Technologies Symposium (PETS), 2021 - 2022
IEEE European Symposium on Security and Privacy (EuroS&P), 2021
ACM CCS Workshop on Artificial Intelligence and Security, 2021
ACM CCS Privacy Preserving Machine Learning Workshop, 2019
IEEE S&P Deep Learning and Security Workshop, 2019 - 2020
Machine Learning and Computer Security Workshop (co-located with NIPS), 2017

Peer reviewer.
Journal of Machine Learning Research (JMLR), 2021 - 2022
International Conference on Learning Representations (ICLR), 2019 - 2023
Neural Information Processing Systems (NeurIPS), 2018 - 2022
International Conference on Machine Learning (ICML), 2018 - 2024
Financial Cryptography and Data Security (FC), 2018
IEEE Symposium on Security & Privacy (IEEE S&P), 2017
Privacy Enhancing Technologies Symposium (PETS), 2016

Outstanding reviewer awards.
International Conference on Machine Learning (ICML), 2020
Neural Information Processing Systems (NeurIPS), 2019
International Conference on Learning Representations (ICLR), 2019

Invited Talks

Symposium on Foundations of Responsible Computing (FORC), Keynote. *Stealing a Generative AI's Secrets (Responsibly), 2024.*
NYU. *Stealing a Generative AI's Secrets (Responsibly)*, 2024.
EPFL Applied Machine Learning Days. *Un-aligning Large Language Models*, 2024.
NeurIPS Workshop on Backdoors in Deep Learning. *Universal jailbreak backdoors from poisoned human feedback*, 2023.
NeurIPS Workshop on Privacy Preserving Federated Learning Document VQA. *Privacy Side-channels in Machine Learning Systems*, 2023.
ICCV Workshop on Out Of Distribution Generalization in Computer Vision. *Is anything really OOD anymore*, 2023.
ICCV Workshop on Adversarial Robustness In the Real World. *Attacking Machine Learning Systems*, 2023.
MLSys Workshop on Decentralized and Collaborative Learning. *Poisoning Web-Scale Training Datasets is Practical*, 2023.
Facebook. *Generative models have the memory of an elephant*, 2023.
Microsoft. *Generative models have the memory of an elephant*, 2023.
AAAI Workshop on Practical Deep Learning in the Wild. *Generative models have the memory of an elephant*, 2023.
Cyber-Defence Campus Conference. *Machine Learning to the Rescue: Risks and Opportunities*, 2022.
Machine Learning Security Seminar Series. *Why you should treat your ML defense like a theorem*, 2022.
Privacy and Security in ML Seminars. *From average-case to worst-case privacy leakage in neural networks*, 2022.
Apple. *What Does it Mean for a Language Model to Preserve Privacy?*, 2022.
AAAI Workshop on Adversarial Machine Learning and Beyond. *When not to use adversarial examples.*, 2022.
KDD Workshop on Adversarial Machine Learning. *Does Adversarial Machine Learning Research Matter?*, 2021.
CVPR Workshop on Media Forensics. *Data poisoning won't save you from facial recognition*, 2021.
Boston-area DP Seminar. *What is (and isn't) Private Learning?*, 2021.
ITASEC Workshop on AI Security. *What is (and isn't) Private Learning?*, 2021.
University of Toronto. *Measuring and Enhancing the Security of Machine Learning*, 2021.
University of Waterloo. *Measuring and Enhancing the Security of Machine Learning*, 2021.
Facebook Research. *Measuring and Enhancing the Security of Machine Learning*, 2021.
Aarhus University. *Measuring and Enhancing the Security of Machine Learning*, 2021.
Google Brain. *Measuring and Enhancing the Security of Machine Learning*, 2021.
ETH Zürich. *Measuring and Enhancing the Security of Machine Learning*, 2021.
CISPA. *Measuring and Enhancing the Security of Machine Learning*, 2021.
Max Plank Institute. *Measuring and Enhancing the Security of Machine Learning*, 2021.
Microsoft Research. *Measuring and Enhancing the Security of Machine Learning*, 2021.
Ruhr University Bochum. *Measuring and Enhancing the Security of Machine Learning*, 2021.
EPFL. *Measuring and Enhancing the Security of Machine Learning*, 2021.
Google. *Differentially Private Learning Needs Better Features* , 2021.
Apple. *Differentially Private Learning Needs Better Features* , 2021.
ECCV CV-COPS Workshop. *Don't use Computer Vision for Web Security*, 2020.

ETH ZISC Seminar. *Developments in Adversarial Machine Learning*, 2019.
Hughes Network Systems. *Defeating Perceptual Ad-Blocking with Adversarial Examples*, 2019.
Stanford Computer Forum. *Defeating Perceptual Ad-Blocking with Adversarial Examples*, 2019.
Palo Alto Networks. *Defeating Perceptual Ad-Blocking with Adversarial Examples*, 2019.
Ad-Blocking Dev Summit. *Defeating Perceptual Ad-Blocking with Adversarial Examples*, 2018.
MIT Bitcoin Expo. *GasToken: A Journey Through Blockchain Resource Arbitrage*, 2018.
Intel. *A Tour of Machine Learning Security*, 2018.
Intel. *Slalom: Fast, Verifiable and Private Execution of Neural Networks in Trusted Hardware*, 2018.
Stanford Innovative Technology Leader program. *Ensemble Adversarial Training*, 2018.
Facebook. *Ensemble Adversarial Training*, 2017.
Cybersecurity with the Best. *Ensemble Adversarial Training*, 2017.
Berkeley Security Seminar. *Ensemble Adversarial Training*, 2017.
MLconf. *FairTest: Discovering Unwarranted Associations in Data-Driven Applications*, 2016.

Selected Press Coverage

Ars Technica	*“Tool preventing AI mimicry cracked; artists wonder what’s next”*	2024
SRF	*“KI-Benchmarks haben mehr Probleme als Lösungen”*	2024
Le Monde	*“The thousand and one ways to derail artificial intelligence”*	2024
The Atlantic	*“The Flaw That Could Ruin Generative AI”*	2024
ZD Net	*“ChatGPT can leak training data, violate privacy”*	2023
Tech Xplore	*“Trick prompts ChatGPT to leak private data”*	2023
Business Insider	*“Google researchers say they got OpenAI’s ChatGPT to reveal some of its training data with just one word”*	2023
RTS	*“Intelligence Artificielle: Ange ou Démon?”*	2023
Science	*“Alarmed tech leaders call for AI research pause”*	2023
The Economist	*“It doesn’t take much to make machine-learning algorithms go awry”*	2023
MIT Tech Review	*“Three ways AI chatbots are a security disaster”*	2023
IEEE Spectrum	*“Protecting AI Models from ”Data Poisoning””*	2023
ZD Net	*“The next big threat to AI might already be lurking on the web”*	2023
The Register	*“It is possible to extract copies of images used to train generative AI models”*	2023
New Scientist	*“AI image generators that create close copies could be a legal headache”*	2023
MIT Tech Review	*“AI models spit out photos of real people and copyrighted images”*	2023
Ars Technica	*“Stable Diffusion memorizes some images, sparking privacy concern”*	2023
Motherboard	*“AI Spits Out Exact Copies of Training Images, Real People, Logos, Researchers Find”*	2023
VentureBeat	*“Is AI moving too fast for ethics?”*	2022
MIT Tech Review	*“What does GPT-3 know about me?”*	2022
Tech Xplore	*“The risks of attacks that involve poisoning training data for machine learning models”*	2022
Wired	*“GitHub’s Commercial AI Tool Was Built From Open Source Code”*	2021
The Register	*“What happens when your massive text-generating neural net starts spitting out people’s phone numbers?”*	2021
Nature	*“Robo-writers: the rise and risks of language-generating AI”*	2021
Wired	*“Even Privacy-Focused Cryptocurrency Can Spill Your Secrets”*	2019
Slashdot	*“Researchers Defeat Perceptual Ad Blockers, Declare New Arms Race”*	2018
Motherboard	*“Researchers Defeat Most Powerful Ad Blockers, Declare a ’New Arms Race”’*	2018
Coindesk	*“Smarter Bug Bounties? Hydra Codes Creative Solution for Ethereum Theft”*	2017
CACM	*“How to Steal the Mind of an AI”*	2016
The Register	*“How to Steal the Mind of an AI”*	2016
Wired	*“How to steal an AI”*	2016
Quartz	*“Stealing an AI algorithm and its underlying data is a high-school level exercise”*	2016