Roy Rinberg

www.royrinberg.com
royrinberg+CV@gmail.com
Github: RoyRin

Selected Work and Research Experience

Harvard University, Cambridge, MA — Aug. 2023 - Present
PhD researcher on AI Security, machine unlearning, fundamentals of Differential Privacy (DP). Advised by Prof. Boaz Barak and Prof. Salil Vadhan.

MATS (ML Alignment and Theory Scholar), Berkeley, CA — Summer 2025
Research Scholar at MATS, an independent research program for research on AI safety. Advised by Keri Warr and Nicholas Carlini at Anthropic on detecting and preventing model weight exfiltration.

Columbia University, New York, NY — Aug. 2021 - Aug. 2023
Master's research on the fundamentals of Differential Privacy and ML. Advised Prof. Rachel Cummings and Prof. Steven Bellovin.

Shelton AI, New York, NY — Jan. 2022 - Jun. 2022
Founding Software Engineer at Shelton AI, a fintech startup that helps pension funds manage investments. I developed core AWS infrastructure for NLP document processing pipeline.

Ouster, San Francisco, CA — Jun. 2018 - Jul. 2021
Early stage software engineer at Ouster, a lidar sensor company. Led on-edge computing development and platforms for evaluating safety algorithms on historical and real-time data.

Selected Papers

Full list: Google Scholar

AI Security

- **R. Rinberg**, A. Karvonen, A. Hoover, D. Reuter, K. Warr. Verifying LLM Inference to Prevent Model Weight Exfiltration. (2025). arXiv preprint.

Machine Unlearning

- **R. Rinberg**, U. Bhalla, I. Shilov, R. Gandikota. RippleBench: Capturing Ripple Effects by Leveraging Existing Knowledge Repositories. (2025). NeurIPS MechInterp Workshop (*Spotlight*).
- **R. Rinberg**, P. Puigdemont, M. Pawelczyk, V. Cevher. Data-Unlearn-Bench: Making Evaluating Data Unlearning Easy. (2025). ICML Machine Unlearning for GenAI Workshop.
- **R. Rinberg**, K. Georgiev, S. Park, S. Garg, A. Ilyas, A. Madry, S. Neel. Attribute-to-Delete: Machine Unlearning via Datamodel Matching. (2024). ICLR 2025.

Differential Privacy

- **R. Rinberg**, Ilia Shumailov, Rachel Cummings, Nicolas Papernot. Beyond Laplace and Gaussian: Exploring the Generalized Gaussian Mechanism for Private Machine Learning. Preprint.
- F. Boenisch, C Mühl, A. Dziedzic, **R. Rinberg**, N. Papernot. Have it your way: Individualized Privacy Assignment for DP-SGD. *Accepted to Neurips 2023*.
- F. Boenisch, C Mühl, **R. Rinberg**, J. Ihrig, A. Dziedzic. Individualized PATE: Differentially Private Machine Learning with Individual Privacy Guarantees. *Accepted to PoPETs 2023*.

Education

Harvard University, Cambridge, MA — 2023 - Present
PhD. Computer Science. Advisors: Prof. Salil Vadhan and Prof. Boaz Barak

Columbia University, New York, NY — 2021 - 2023
MS in Computer Science [Thesis Track]. Advisors: Prof. Rachel Cummings and Prof. Steven Bellovin

New York University, New York, NY — 2014 - 2018
B.A. Computer Science, Physics, Minor: Math.

Teaching and Service

Teaching: TF for CS2881 (AI Safety, Fall '25); Head TF for CS1200 (Intro to Algorithms, Fall '24); Physics I/II Tutor at NYU ('17-'18)

Organizing: Founding Organizer: Technically Private - group of graduate students who work on privacy and security ('21-Present); Co-founder, Project BEST - Education non-profit, Fundraised and grew organization to 25 chapters across 3 states, reaching 3000+ students. ('11-'14)

Community Service: Mentor at Mentor Ukraine for 3 students ('22-'23). Advocated Ouster to set up recurring donations to public-interest orgs and paid volunteer days ('18-'20)

Academic Service: Reviewer for NeurIPS ('23, '24, '25), ICML ('23, '25), ICLR ('23, '24); Assistant organizer for OSDI '23 PC