

Roy Rinberg

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

SELECTED WORK AND RESEARCH EXPERIENCE	Harvard University, Cambridge, MA	AUG. 2023 - PRESENT
	<i>PhD researcher on AI Security, machine unlearning, fundamentals of Differential Privacy (DP). Advised by Prof. Boaz Barak and Prof. Salil Vadhan.</i>	
	MATS (ML Alignment and Theory Scholar), Berkeley, CA	SUMMER 2025
	<i>Research Scholar at MATS, an independent research program for research on AI safety. Worked with Keri Warr and Nicholas Carlini at Anthropic on detecting and preventing model weight exfiltration.</i>	
	Columbia University, New York, NY	AUG. 2021 - AUG. 2023
	<i>Master's research on the fundamentals of Differential Privacy and machine learning.</i>	
	Shelton AI, New York, NY	JAN. 2022 - JUN. 2022
	<i>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.</i>	
	Ouster, San Francisco, CA	JUN. 2018 - JUL. 2021
	<i>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.</i>	

SELECTED PAPERS Full list: Google Scholar	AI Security
	<ul style="list-style-type: none">• R. Rinberg, A. Karvonen, A. Hoover, D. Reuter, K. Warr. <u>Verifying LLM Inference to Prevent Model Weight Exfiltration</u>. (2025). arXiv preprint.
	Machine Unlearning
	<ul style="list-style-type: none">• R. Rinberg, U. Bhalla, I. Shilov, R. Gandikota. <u>RippleBench: Capturing Ripple Effects by Leveraging Existing Knowledge Repositories</u>. (2025). NeurIPS MechInterp Workshop (<i>Spotlight</i>).• R. Rinberg, P. Puigdemont, M. Pawelczyk, V. Cevher. <u>Data-Unlearn-Bench: Making Evaluating Data Unlearning Easy</u>. (2025). ICML Machine Unlearning for GenAI Workshop.• R. Rinberg, K. Georgiev, S. Park, S. Garg, A. Ilyas, A. Madry, S. Neel. <u>Attribute-to-Delete: Machine Unlearning via Datamodel Matching</u>. (2024). ICLR 2025.
	Differential Privacy
	<ul style="list-style-type: none">• R. Rinberg, Ilia Shumailov, Rachel Cummings, Nicolas Papernot. <u>Beyond Laplace and Gaussian: Exploring the Generalized Gaussian Mechanism for Private Machine Learning</u>. Preprint.• F. Boenisch, C Mühl, A. Dziedzic, R. Rinberg, N. Papernot. <u>Have it your way: Individualized Privacy Assignment for DP-SGD</u>. <i>Accepted to Neurips 2023</i>.• F. Boenisch, C Mühl, R. Rinberg, J. Ihrig, A. Dziedzic. <u>Individualized PATE: Differentially Private Machine Learning with Individual Privacy Guarantees</u>. <i>Accepted to PoPETs 2023</i>.

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	