

Rishi D. Jha

PHD STUDENT · SECURITY RESEARCHER

☎ 425.677.4846 | ✉ rdj58@cornell.edu | 🏠 rishijha.com | 📱 rjha18 | 🌐 rishi-jha

Education

Cornell Tech

PHD STUDENT, COMPUTER SCIENCE

New York / Ithaca, NY

Aug. 2023 - Present

- Advisor: Prof. Vitaly Shmatikov
- Funded by the Cornell University Fellowship (20% of incoming PhDs) for my first year

University of Washington — Seattle

MS., COMPUTER SCIENCE

Seattle, WA

Sep. 2022 - Jun. 2023

- Master's Thesis: *Label Poisoning is All You Need*
- Advisor: Prof. Sewoong Oh

University of Washington — Seattle

BS.BA., COMPUTER SCIENCE AND MATHEMATICS — PHILOSOPHY: *Cum Laude, Phi Beta Kappa*

Seattle, WA

Sep. 2018 - Mar. 2022

- Jun. 2022: Graduated Cum Laude with Phi Beta Kappa honors
- 2018-22: Dean's List (all eligible quarters)
- GPA: 3.84 / 4.0

Awards and Honors

2024	Distinguished Paper Award , USENIX Security — Top 22 papers (out of 417)	Philadelphia, PA
2024	Graduate Research Fellowship Program Honorable Mention , NSF	USA
2023	Cornell University Fellowship , Cornell University — 20% of incoming PhDs	Ithaca, NY
2022	Phi Beta Kappa , University of Washington	Seattle, WA
2022	Cum Laude , University of Washington — Top 10% across Arts & Sciences	Seattle, WA
2018-22	Dean's List , University of Washington — All eligible quarters	Seattle, WA
2021-22	Varsity Climbing Team , University of Washington	Seattle, WA
2019	Finalist , (Top 4 of 36 Teams) UW Foster CBDC: Consulting Challenge	Seattle, WA
2018	National Merit Finalist , National Merit Scholarship	USA
2017	3rd Place , Microsoft OneWeek Hackathon Consumer Category — 1000+ Teams	Redmond, WA

Publications

CONFERENCE

- [1] Tingwei Zhang*, **Rishi Jha***, Eugene Bagdasaryan, and Vitaly Shmatikov. "Adversarial Illusions in Multi-Modal Embeddings". In: *33rd USENIX Security Symposium (USENIX)*. Received the **Distinguished Paper Award** (5% of accepted papers). Aug. 2024.
- [2] **Rishi Jha***, Jonathan Hayase*, and Sewoong Oh. "Label Poisoning is All You Need". In: *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*. Dec. 2023.
- [3] Dimitrios C. Gklezakos, **Rishi Jha**, and Rajesh P.N. Rao. "Hyper-Universal Policy Approximation: Learning to Generate Actions from a Single Image using Hypernets". In: *Neurovision 2022: A CVPR Workshop (Neurovision @ CVPR)*. June 2022.
- [4] **Rishi Jha** and Kai Mihata. "On Geodesic Distances and Contextual Embedding Compression for Text Classification". In: *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15 @ NAACL)*. June 2021.

MASTER'S THESIS

- [5] **Rishi Jha**. "Label Poisoning is All You Need". University of Washington, Seattle, 2023.

PATENTS (PENDING)

- [6] Nisha S. Hameed, **Rishi D. Jha**, and Evan Argyle. "Graph-Based Analysis of Security Incidents". U.S. pat. Microsoft. 2022.

Academic Research

Cornell Security

PHD STUDENT

New York, NY

Aug 2023 – Present

Working with **Prof. Vitaly Shmatikov** to make the machine learning we rely on more robust, secure, and private. Currently exploring how the geometry of language embeddings can be exploited to reveal (potentially-sensitive) information.

- Previously, we showed that commercial multi-modal embeddings are vulnerable to "adversarial illusions" in which an adversary perturbs a given image or sound to make its embedding close to an arbitrary input in another modality. These illusions exploit proximity in the embedding space and are thus agnostic to downstream tasks and modalities, enabling a wholesale compromise of current and future tasks, as well as modalities not available to the adversary. **Distinguished Paper** at **USENIX Security 2024** [1].

Sewoong Lab — Foundations of Machine Learning

GRADUATE RESEARCH ASSISTANT

Seattle, WA

May 2021 – Aug 2023

Worked with **Prof. Sewoong Oh** and **Jonathan Hayase** to:

- (Master's Thesis Project) Develop a novel trajectory-matching-based backdoor attack, FLIP, that corrupts (i.e., 'poisons') only the labels in a training set to create a backdoor with an arbitrary trigger. In particular, we show that with few-shot poisons (i.e., less than 1% of a dataset's training labels), FLIP can inject a backdoor with a 99.6% success rate while remaining undetected with less than a 1% degradation of clean accuracy. We also demonstrate FLIP's surprising robustness to dataset, trigger, and architecture. Thesis submitted in **June 2023** [5]. Paper accepted at **NeurIPS 2023** [2].
- Create an open-source 'backdoor'-attack-benchmark platform and survey. Code can be found [here](#).

Center for Neurotechnology

UNDERGRADUATE ML RESEARCHER

Seattle, WA

Mar. 2020 – Aug. 2022

Worked with **Prof. Rajesh Rao** and **Dimitrios Gklezakos** to:

- Develop a low-cost, 'personalized' hypernetwork for hierarchical and task-conditional RL called the Hyper-Universal Policy Approximator (HUPA). HUPAs are up to 35% more resilient to sparsity and have up to 25% better generalization than their traditional embedding alternatives. Paper accepted at **NeuroVision '22 at CVPR** [3].
- Construct an audio-visual hypernetwork for representation learning and classification on a massive dataset in which a video-controlled neural network controls the weights of an audio interpreter.
- Create a convolutional, manifold-learning based network to learn complex features in natural images in an unsupervised fashion using sparse coding. The system learns representational similarities between features and generalizes them.

Self-Directed

NLP RESEARCHER

Seattle, WA

Nov. 2020 – Jun. 2021

Paper accepted at **TextGraphs '21 at NAACL** [4]. Worked with **Kai Mihata** to:

- Investigate the downstream effects of compressing BERT embeddings using nonlinear dimensionality reduction techniques and geodesic estimations.
- Find that nonlinear compressions of the embeddings tend to work well in some data regimes, a feature that can be utilized in memory-constrained settings.

ICTD Lab

UNDERGRADUATE RESEARCHER

Seattle, WA

Nov. 2018 – May 2019

Worked with **Spencer Sevilla** to:

- Investigate the performance dynamics of different chat apps in poor network conditions.
- Implement a teaching solution for schoolchildren in rural Indonesia.

Research in Industry

Microsoft Defender Research

SOFTWARE ENGINEERING INTERN — DATA SCIENCE

Redmond, WA

Jul. 2022 – Sep. 2022

- Ideated, pitched, and implemented a low-cost, humanly interpretable meta-learning framework that exploits spectral similarities in existing classifier responses to drive robustness in the Defender product. The productionalized system was lightweight, had upwards of 97% precision and recall, and was humanly interpretable.
- After my departure, the model was being pushed from pre-production to production with the goal of providing protection for billions of users by the **end of 2023**.

Microsoft Defender Research

SOFTWARE ENGINEERING INTERN — DATA SCIENCE

Remote

Jun. 2021 – Sep. 2021

Patent submitted in **Winter 2022** [6].

- Ideated and designed patent-pending approach to detect malicious Command-and-Control intrusions in corporate networks using spectral methods on graphs. The model achieved high precision and recall in finding Indicators of Compromise in historical data.
- The project received significant investment from the team and Microsoft Research (MSR) since my departure with a goal of pushing an extension of the model to production in **Summer 2023**.

Teaching

Cornell Tech

TA FOR "PRIVACY IN THE DIGITAL AGE"

New York, NY

Aug. 2024 - Dec. 2024

Currently TAing 60+ students in the graduate-level privacy course with **Profs. Vitaly Shmatikov** and **Helen Nissenbaum**.

University of Washington — Seattle

Seattle, WA

4X UNDERGRAD / GRAD MACHINE LEARNING TA

Mar. 2020 - Dec. 2021

During Spring 2020, Winter 2021, Spring 2021, Autumn 2021:

- Taught undergraduate and graduate students as an undergraduate through 25-person sections and biweekly office hours.
- Designed section materials for entire teaching staff, monitored discussion boards, and graded assignments.

University of Washington — Seattle

Seattle, WA

MACHINE LEARNING COURSE DESIGNER

Jun. 2021 - Sep. 2021

During Summer 2021, funded by **Prof. Sewoong Oh** to:

- Redesign the course's problem sets and homework infrastructure to keep up with a rapidly evolving course and field, and lower the barrier of entry to machine learning.
- Drive equitability by adding necessary data context, removing technical jargon, and constructing homework problems that required students to challenge algorithmic and implicit biases in machine learning.
- Create a new central grading system and TA codebase for future quarters and course staffs to use.

Other Work Experience

Microsoft

Remote

SOFTWARE ENGINEERING INTERN — DEFENDER SECURITY

Jun. 2020 - Sep. 2020

- Reduced related COGS by \$100K - \$1M by creating ML model to selectively download dangerous files for analysis. In production.
- Built infrastructure for safer ML model deployment. In production.
- Decreased researcher rule development time by 35%, by creating VSCode extension to natively test rules. In production.

Microsoft

Redmond, WA

EXPLORE INTERN — OFFICE.COM FRONT END

Jun. 2019 - Aug. 2019

- Designed, implemented, and released front end notes tool for the Office.com team using Typescript, Redux, and React internally.

Skills

Interests Machine Learning, NLP, Security, Privacy, Embeddings

Technical Python, PyTorch, HuggingFace, TensorFlow, JAX, C++, Java / C#,

Languages English, Hindi, Spanish

Service

2024 **Reviewer**, ICLR

Remote

2023 **Reviewer**, ICLR

Remote

2023 **Reviewer**, ICML

Remote

2021 **Presenter**, High School Neuroscience Club @ The Overlake School

Redmond, WA

Selected Coursework

Machine Learning Machine Learning[†], NLP[†], Deep Learning Theory[†], Reinforcement Learning[†], Deep Learning

Other Computer Science Cryptography[†], Privacy[†], Security[†], Human-Centered AI[†], Algorithms, Databases

Mathematics Real Analysis I & II, Probability and Statistics I, II, & III, Modern Algebra I & II, Linear Algebra

[†]Taken at both the undergraduate and PhD levels.

[†]Taken at the PhD level.