33 Oxford Street Rm 110
Cambridge MA 02138

ARIEL HERBERT-VOSS

ariel_herbertvoss@g.harvard.edu
@adversariel

EDUCATION

Harvard University, Cambridge, MA — 2016 - present
Ph.D. Computer Science — *Courses*: Multi-Agent Intelligent Systems, Computer Security, Machine Learning, Distributed Systems, Differential Privacy, Economics of National Security

University of Utah, Salt Lake City, UT — 2011 - 2016
BSc. Computer Science, *Honors* — *Courses*: Artificial Intelligence, Computer Vision, Data Mining, Algorithms, Data Visualization, Embedded Systems, Software Engineering
BSc. Mathematics, *Pi Mu Epsilon* — *Courses*: Matrix Theory, Numerical Analysis, Probability Theory, Differential Equations, Linear Algebra, Real Analysis

RESEARCH EXPERIENCE

OpenAI — Research Affiliate — 2019-present
- Building adversary threat models to inform researchers, business leaders, and policymakers about security implications of AI research.

Vector Institute for Artificial Intelligence — Graduate Student Affiliate — 2018-present
- Ongoing research on adversarial machine learning.

Rapid7 — Data Science Student Co-op — 2018
- Developed machine learning tools for analyzing and identifying network anomalies.

MIT Lincoln Laboratory — Summer Technical Intern — 2016
- Developed neural network model using cochleagram data representations for speech formant tracking to make predictions about speech-in-noise performance for hearing loss.

Cornell University Department of Mathematics — NSF REU Program — 2013
- Developed smooth interpolation algorithm for high-dimensional data. Created a MATLAB package.

PUBLICATIONS

Computing Minimal Interpolants in $C^{1,1}(\mathbb{R}^d)$
Herbert-Voss, A., Hirn, M.J., McCollum, F. *Revista Matemática Iberoamericana*, vol. 33, pg 29-66. 2017.

Discovering Human Cortical Organization using Data-Driven Analyses of fMRI
Liu, T., Herbert-Voss, A. *Center for Brains Minds and Machines Memo No. 024*, pg 31. 2014.

PRESENTATIONS

Machine Learning Model Hardening for Fun and Profit — 2018
DEF CON 26 AI Village
Caesar's Palace, Las Vegas, NV. Aug 11

Model Inversion Attacks and Differential Privacy Defenses — 2018
DEF CON China [Beta]
Kuntai Hotel, Beijing, China. May 12

Smooth Interpolants for Manifold Learning — 2015
8th International Whitney Problems Workshop
Centre International de Rencontres Mathématiques (CIRM), Marseille, France. Oct 19-23

HONORS and AWARDS

Belfer Center Cybersecurity Fellowship — 2019-2020
Berkman-Klein Center for Internet & Society Fellowship — 2019-2020
Harvard Graduate Student Leadership Institute Fellowship — 2016
MIT Lincoln Laboratory I3 Challenge Grand Prize — 2016
Google Lime Scholarship — 2015-2016
Utah Flagship Scholarship — 2011-2015

ADDITIONAL EXPERIENCE

DEF CON AI Village — Co-founder, Board Member — 2018
- Co-organize bi-annual workshop for the hacker community on use and misuse of AI technology.

Summary

My primary research focus is on adversarial machine learning and security. To address the gap between proposed attacks and real security threats to machine learning systems, I replicated state-of-the-art attacks from the academic literature in order to understand the conditions needed to replicate these attacks in real adversarial scenarios. I found that not only are the proposed attacks quite brittle when deployed in realistic environments, but in almost all cases there are simpler ways to accomplish attacker goals. For example, I found that Tesla's autonomous lane-finding algorithm is not robust to simple environmental changes such as adding road salt, indicating that a real adversary need not go through the effort of generating adversarial examples to cause the car to crash. This indicates a greater need for building robustness at the systems level in addition to the algorithmic level which is the current focus in the AI research community.

Research on the adversarial dynamics governing vulnerability discovery and remediation indicates that vulnerabilities are exploited by virtue of the adversarial environment, not inherent features of the code. I am developing game theoretical models to capture this in the context of machine learning model vulnerability to data poisoning and adversarial example attacks. During this project, I've encountered relevant economics research on modeling adversary incentives. I am also interested in learning how to incentivize research focused on real threats to machine learning systems from a policy level. I am eager to find collaborators in economics with shared interests in AI and security.