

# Yigitcan Kaya

(+1) 240 416 7939 | yigitcankaya94@gmail.com | <https://yigitcankaya.github.io>

## Research Interests

Trustworthy Artificial Intelligence | Artificial Intelligence for Security Applications | Data-Driven Security

## Education

**Ph.D. in Computer Science, University of Maryland – College Park, MD** 2017–2023  
**B.Sc. in Computer Engineering, Bilkent University – Ankara, Turkey** 2012–2017

## Honors & Awards

- **Fellow**, US Intelligence Community (IC) Postdoctoral Fellowship Program 2023–Present
- **Travel Scholarship**, IEEE SaTML 2025
- **Fellow**, University of Maryland, Clark School Future Faculty Program 2022–2023
- **Honorable Mention**, NSF Graduate Research Fellowship (GRFP) 2019
- **Dean’s Fellowship**, University of Maryland 2017–2018
- **Comprehensive Scholarship**, Bilkent University (full tuition & stipend) 2012–2017

## Professional Experience

**UC Santa Barbara, SecLab — Postdoctoral Fellow** Oct 2023–Present

- Mentored summer interns and junior lab members on research design, reproducible experimentation, and paper submissions.
- Researched realistic evaluations of ML methods for security applications; identifying and addressing reliability challenges.
- Developed a retrieval-augmented generation (RAG) component for Team Shellphish’s AI-powered vulnerability patching system that advanced to the DARPA AIXCC finals.

**Amazon Web Services, AI Research Team — Applied Scientist Intern** Jun 2021–May 2022

- Built ML security and robustness solutions for AWS customers.
- Published an ICML 2022 paper [P-8] proposing an advanced attack against adversarial example detectors.

**Amazon Web Services, Identity Team — Applied Scientist Intern** Sep 2020–Dec 2020

- Created an explainability solution for Haze1, an ML system that assists AWS customers with access management.

## Funding & Proposal Development

*Authored and co-led multiple competitive proposals funding my Ph.D. and postdoctoral research through federal fellowships and industry awards, consistently aligning AI security initiatives with agency priorities and collaborative research efforts.*

**[F-5] RedactBench: A Formal Framework for LLM-based Confidential Information Redaction** 2025  
US IC Postdoctoral Fellowship — Third-Year Extension. *Sole author.* Developed proposal in collaboration with U.S. Government sponsors; awarded ~\$95,000.

**[F-4] Combating False Positives in ML-Based Security Applications with Context-Adaptive Classification** 2024  
Amazon Research Awards. *Led proposal writing.* Drafted and coordinated the submission; secured \$80,000 in research funding and \$20,000 in cloud credits for the lab.

**[F-3] Adaptable Machine Learning Systems for Antifragile Cyber Defenses** 2023  
US IC Postdoctoral Fellowship. *Sole author.* Built upon prior work [P-11], [P-9], and [P-8]. Fellowship fully funded my postdoctoral research at UCSB for two years; awarded ~\$190,000.

**[F-2] Distinct Impact of Trojans on the Internal Behaviors of Deep Neural Networks** 2019  
IARPA, Trojans in Artificial Intelligence (TrojAI). *Co-authored with Ph.D. advisor.* Contributed to proposal development and experimental design based on findings in [P-2], enabling the lab’s participation in IARPA’s TrojAI program.

## [F-1] Functional and Semantic Understanding of Deep Learning

2018

Laboratory for Telecommunication Sciences (LTS). *Co-authored with Ph.D. advisor.* Extended prior research from [P-2]; helped secure funding that supported my Ph.D. research.

## Conference Publications

---

*13 peer-reviewed papers in leading AI and Security venues (IEEE S&P, USENIX Security, ICLR, ICML, NeurIPS, SaTML), including oral and spotlight presentations on trustworthy AI systems.*

- [P-13] “When AI Meets the Web: Prompt Injection Risks in Third-Party AI Chatbot Plugins”  
**Yigitcan Kaya**, Anton Landerer, Stijn Pletinckx, Michelle Zimmermann, Christopher Kruegel, Giovanni Vigna; **IEEE S&P 2026**
- [P-12] “PoisonedParrot: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models”  
Michael-Andrei Panaitecu-Liess, Pankayaraj Pathmanathan, **Yigitcan Kaya**, Zora Che, Bang An, Sicheng Zhu, Aakriti Agrawal, Furong Huang; **NAACL 2025** (Oral presentation)
- [P-11] “ML-Based Behavioral Malware Detection Is Far From a Solved Problem”  
**Yigitcan Kaya**, Yizheng Chen, Marcus Botacin, Shoumik Saha, Fabio Pierazzi, Lorenzo Cavallaro, David Wagner, Tudor Dumitras; **SaTML 2025**
- [P-10] “Like Oil and Water: Group Robustness Methods and Poisoning Defenses Don’t Mix”  
Michael-Andrei Panaitecu-Liess\*, **Yigitcan Kaya**\*, Sicheng Zhu, Furong Huang, Tudor Dumitras; **ICLR 2024**
- [P-9] “DRSM: De-Randomized Smoothing on Malware Classifier Providing Certified Robustness”  
Shoumik Saha, Wenxiao Wang, **Yigitcan Kaya**, Soheil Feizi, Tudor Dumitras; **ICLR 2024**
- [P-8] “Generating Distributional Adversarial Examples to Evade Statistical Detectors”  
**Yigitcan Kaya**, Bilal Zafar, Sergul Aydore, Nathalie Rauschmayr, Krishnam Kenthapadi; **ICML 2022**
- [P-7] “Qu-ANTI-zation: Exploiting Quantization Artifacts for Achieving Adversarial Outcomes”  
Sanghyun Hong, Michael-Andrei Panaitecu-Liess, **Yigitcan Kaya**, Tudor Dumitras; **NeurIPS 2021**
- [P-6] “When Does Data Augmentation Help With Membership Inference Attacks?”  
**Yigitcan Kaya**, Tudor Dumitras; **ICML 2021**
- [P-5] “A Panda? No, It’s a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Network Inference”  
Sanghyun Hong\*, **Yigitcan Kaya**\*, Ionuț-Vlad Modoranu, Tudor Dumitras; **ICLR 2021** (Spotlight presentation)
- [P-4] “How to Own the NAS in Your Spare Time”  
Sanghyun Hong, Michael Davinroy, **Yigitcan Kaya**, Dana Dachman-Soled, Tudor Dumitras; **ICLR 2020**
- [P-3] “Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks”  
Sanghyun Hong, Pietro Frigo, **Yigitcan Kaya**, Cristiano Giuffrida, Tudor Dumitras; **USENIX Security 2019**
- [P-2] “Shallow-Deep Networks: Understanding and Mitigating Network Overthinking”  
**Yigitcan Kaya**, Sanghyun Hong, Tudor Dumitras; **ICML 2019**
- [P-1] “When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks”  
Octavian Suci, Radu Marginean, **Yigitcan Kaya**, Hal Daumé, Tudor Dumitras; **USENIX Security 2018**

## Workshop Publications

---

*Collaborations with interns and junior researchers on emerging topics in AI security and interpretability.*

- [W-3] “Demystifying Cipher-Following in Large Language Models via Activation Analysis”  
Megan Gross, **Yigitcan Kaya**, Christopher Kruegel, Giovanni Vigna; **Mechanistic Interpretability Workshop at NeurIPS 2025**
- [W-2] “MADCAT: Combating Malware Detection Under Concept Drift with Test-Time Adaptation”  
Eunjin Roh, **Yigitcan Kaya**, Christopher Kruegel, Giovanni Vigna, Sanghyun Hong; **Workshop on Test-Time Adaptation at ICML 2025**
- [W-1] “Too Big to FAIL: What You Need to Know Before Attacking a Machine Learning System”  
Tudor Dumitras, **Yigitcan Kaya**, Radu Mărginean, Octavian Suci; **International Workshop on Security Protocols 2018**

## Media Coverage

---

*Featured by MIT Technology Review, VentureBeat, ORISE, and other outlets for research advancing AI security and robustness.*

- [M-5] **“Computer Scientist researches the complexity of artificial intelligence for its usage against cyber-attacks”**, Oak Ridge Institute for Science and Education (ORISE), 2024.  
<https://orise.orau.gov/people/success-stories/2024/yigitcan-kaya.html>  
*Featured profile highlighting my IC Postdoctoral Fellowship and research on antifragile AI defenses for cybersecurity.*
- [M-4] Ben Dickson, **“Machine learning security needs new perspectives and incentives”**, VentureBeat, June 2021.  
<https://venturebeat.com/ai/machine-learning-security-needs-new-perspectives-and-incentives>  
*Quoted in a feature discussing systemic challenges in ML security research and incentives for building safer AI systems.*
- [M-3] Ben Dickson, **“Machine learning security: Why protecting neural networks is harder than you think”**, TechTalks, June 2021.  
<https://bdttechtalks.com/2021/06/03/machine-learning-security-neural-networks/>  
*Interview and commentary on my work exposing vulnerabilities in deep learning systems and developing practical defenses.*
- [M-2] Karen Hao, **“This AI could hack your energy bill”**, MIT Technology Review, May 2021.  
<https://www.technologyreview.com/2021/05/06/1024654/ai-energy-hack-adversarial-attack/>  
*Coverage of my research on adversarial machine learning and how AI models can be manipulated to disrupt real-world systems.*
- [M-1] **“Apple’s AirTag security concerns, a deep neural network hack, and an oil pipeline cyberattack”**, DEV News Podcast, Season 4, Episode 5, May 2021.  
<https://dev.to/devnews/s4-e5-apple-s-airtag-security-concerns-a-deep-neural-network-hack-an-oil-pipeline-cyber-attack-and-a-shortage-of-semiconductors>  
*Podcast discussion featuring my work on neural network security and adversarial attacks.*

## Service

---

### Academic

- **Conference Program Committee:** IEEE S&P’25, ’26; CCS’26, USENIX Security’24; SaTML’25, 26, ACSAC’24; RAID’24, ’25
- **Workshop Program Committee:** Dynamic Neural Networks (ICML’22); AdvML Frontiers (ICML’22); Security & Privacy of ML (ICML’19); Adversarial ML in Real-World CV Systems (CVPR’19); Security in ML (NeurIPS’18)
- **Reviewer:** ICML ’20–’24; NeurIPS ’20–’23; ICLR ’22–’24

### Outreach & Teaching

- (2025) Delivered invited mini-lectures on AI safety and societal challenges to community-college students and faculty members.
- (2022) Delivered two invited 2-hour mini-lectures to UMD CS graduate students on security and privacy in machine learning.
- (2020) Delivered a mini-lecture series to Turkish undergraduates on ML research; mentored US graduate school applications.
- (2018) Organized a weekly reading group in the Maryland Cybersecurity Center; nearly doubled participation over prior years.

### Student Mentorship & Supervision

- (2025) Supervised a UCSB summer intern on adversarial risks in RL-based LLM fine-tuning; our joint proposal earned a UCSB Summer Undergraduate Research Fellowship grant (\$4,000).
- (2025) Supervised seven research interns under the NSF ACTION Institute on three projects addressing security and privacy challenges in LLMs; one project formed the basis of [F-5], and another led to the NeurIPS 2025 workshop paper [W-3].
- (2024) Supervised five research interns under the NSF ACTION Institute on security vulnerabilities of AI chatbots on the web; project led to publication [P-13].
- (2021) Supervised two summer research interns on ML-for-security projects; both were admitted to a top US graduate school.
- (2019–2020) Co-advised five summer interns on deep learning security & privacy; work led to publications [P-4], [P-5] and [P-7].

## Talks

---

- [T-11] **UCSB - AI Meetup Series** (Invited Speaker), May 2025  
*When AI Meets the Web: Prompt Injection Risks in Third-Party AI Chatbot Plugins*
- [T-10] **SaTML** (Conference Presentation), April 2025  
*ML-Based Behavioral Malware Detection Is Far From a Solved Problem*

- [T-9] **Visa Research** (Invited Speaker), October 2024  
*Wild Chatbots: A Large-Scale Study of the Trends and Security Flaws in the AI Chatbot Plugin Ecosystem on the Web*
- [T-8] **Intelligence Community Tech Week 2024** (Invited Speaker), September 2024  
*Anti-fragility in Machine Learning-Based Cyber Defenses*
- [T-7] **Chicago Workshop on Coding and Learning** (Invited Speaker), December 2022  
*Wonders and Dangers of Input-Adaptive Neural Network Inference*
- [T-6] **University of Maryland** (Guest Lecturer), November 2022  
Machine Learning Security & Machine Learning Privacy [Host: Dave Levin]
- [T-5] **ICML** (Conference Presentation), July 2022  
*Generating Distributional Adversarial Examples to Evade Statistical Detectors*
- [T-4] **Amazon Web Services Themis Team** (Guest Speaker), November 2021  
*Detecting Adversarial Input Distributions via Layer-wise Statistics*
- [T-3] **Amazon Web Services Science Tech Presentations** (Guest Speaker), July 2021  
*Wonders and Dangers of Input-Adaptive Neural Network Inference*
- [T-2] **ICML** (Conference Presentation), July 2021  
*When Does Data Augmentation Help With Membership Inference Attacks?*
- [T-1] **ICML** (Conference Presentation), July 2019  
*Shallow-Deep Networks: Understanding and Mitigating Network Overthinking*

## Academic References

---

- **Giovanni Vigna**, Professor, CS Department, University of California, Santa Barbara  
vigna@ucsb.edu
- **Christopher Kruegel**, Professor, CS Department, University of California, Santa Barbara  
chris@cs.ucsb.edu
- **David Wagner**, Professor, EECS Department, University of California, Berkeley  
daw@cs.berkeley.edu
- **Lorenzo Cavallaro**, Professor, CS Department, University College London  
l.cavallaro@ucl.ac.uk
- **Tudor Dumitras**, Associate Professor, ECE Department, University of Maryland  
tudor@umd.edu