Yigitcan Kaya

(+1) XXX XXX XXXX | yigitcan@ucsb.edu | https://yigitcankaya.github.io

Research Interests

Trustworthy Artificial Intelligence | Artificial Intelligence for Security Applications | Systems 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 as one of the top seven teams.

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 **[P8]** 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 `Hazel`, 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.

[F5] 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**.

[F4] 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 **$40,000** in cloud credits for the lab.

[F3] Adaptable Machine Learning Systems for Antifragile Cyber Defenses *2023*
US IC Postdoctoral Fellowship. *Sole author.* Built upon prior work **[R2]**, **[P9]**, and **[P8]**. Fellowship fully funded my postdoctoral research at UCSB for two years; awarded ∼**$190,000**.

[F2] 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 **[P2]**, enabling the lab's participation in IARPA's TrojAI program.

[F1] Functional and Semantic Understanding of Deep Learning *2018*
Laboratory for Telecommunication Sciences (LTS). *Co-authored with Ph.D. advisor.* Extended prior research from **[P2]**; helped secure funding that supported my Ph.D. research.

Top-Tier Conference Publications

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

[P11] "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**

[P10] "Like Oil and Water: Group Robustness Methods and Poisoning Defenses Don't Mix"
Michael-Andrei Panaitescu-Liess*, **Yigitcan Kaya***, Sicheng Zhu, Furong Huang, Tudor Dumitras; **ICLR 2024**

[P9] "DRSM: De-Randomized Smoothing on Malware Classifier Providing Certified Robustness"
Shoumik Saha, Wenxiao Wang, **Yigitcan Kaya**, Soheil Feizi, Tudor Dumitras; **ICLR 2024**

[P8] "Generating Distributional Adversarial Examples to Evade Statistical Detectors"
Yigitcan Kaya, Bilal Zafar, Sergul Aydore, Nathalie Rauschmayr, Krishnaram Kenthapadi; **ICML 2022**

[P7] "Qu-ANTI-zation: Exploiting Quantization Artifacts for Achieving Adversarial Outcomes"
Sanghyun Hong, Michael-Andrei Panaitescu-Liess, **Yigitcan Kaya**, Tudor Dumitras; **NeurIPS 2021**

[P6] "When Does Data Augmentation Help With Membership Inference Attacks?"
Yigitcan Kaya, Tudor Dumitras; **ICML 2021**

[P5] "A Panda? No, It's a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Network Inference"
Sanghyun Hong*, **Yigitcan Kaya***, Ionuţ-Vlad Modoranu, Tudor Dumitras; **ICLR 2021** (Spotlight presentation)

[P4] "How to 0wn the NAS in Your Spare Time"
Sanghyun Hong, Michael Davinroy, **Yigitcan Kaya**, Dana Dachman-Soled, Tudor Dumitras; **ICLR 2020**

[P3] "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**

[P2] "Shallow-Deep Networks: Understanding and Mitigating Network Overthinking"
Yigitcan Kaya, Sanghyun Hong, Tudor Dumitras; **ICML 2019**

[P1] "When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks"
Octavian Suciu, Radu Marginean, **Yigitcan Kaya**, Hal Daumé, Tudor Dumitras; **USENIX Security 2018**

Other Conference Publications

Three peer-reviewed papers in respected AI and Security venues, including SaTML and NAACL.

[R3] "PoisonedParrot: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models"
Michael-Andrei Panaitescu-Liess, Pankayaraj Pathmanathan, **Yigitcan Kaya**, Zora Che, Bang An, Sicheng Zhu, Aakriti Agrawal, Furong Huang; **NAACL 2025** (Oral presentation)

[R2] "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**

[R1] "Understanding, uncovering, and mitigating the causes of inference slowdown for language models"
Kamala Varma, Arda Numanoglu, **Yigitcan Kaya**, Tudor Dumitras; **SaTML 2024**

Workshop Publications

Peer-reviewed workshops showcasing collaborations with interns and junior researchers on emerging topics in AI security.

[W3] "Demystifying Cipher-Following in Large Language Models via Activation Analysis"
Megan Gross, **Yigitcan Kaya**, Christopher Kruegel, Giovanni Vigna; **Mechanistic Interpretability Workshop at NeurIPS 2025**

[W2] "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**

[W1] "Too Big to FAIL: What You Need to Know Before Attacking a Machine Learning System"
Tudor Dumitras, **Yigitcan Kaya**, Radu Mărginean, Octavian Suciu; **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.

[M5] **"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.

[M4] 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.

[M3] Ben Dickson, **"Machine learning security: Why protecting neural networks is harder than you think"**, *TechTalks*, June 2021.
https://bdtechtalks.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.

[M2] 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.

[M1] **"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 **[F5]**, and another led to the NeurIPS 2025 workshop paper **[W3]**.
- (2024) Supervised five research interns under the NSF ACTION Institute on security vulnerabilities of AI chatbots on the web; project led to publication **[P11]**.
- (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 **[P4]**, **[P5]** and **[P7]**.

Talks

[T11] **UCSB - AI Meetup Series** (Invited Speaker), May 2025
When AI Meets the Web: Prompt Injection Risks in Third-Party AI Chatbot Plugins

[T10] **SaTML** (Conference Presentation), April 2025
ML-Based Behavioral Malware Detection Is Far From a Solved Problem

[T9] **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

[T8] **Intelligence Community Tech Week 2024** (Invited Speaker), September 2024
Anti-fragility in Machine Learning-Based Cyber Defenses

[T7] **Chicago Workshop on Coding and Learning** (Invited Speaker), December 2022
Wonders and Dangers of Input-Adaptive Neural Network Inference

[T6] **University of Maryland** (Guest Lecturer), November 2022
Machine Learning Security & Machine Learning Privacy [Host: Dave Levin]

[T5] **ICML** (Conference Presentation), July 2022
Generating Distributional Adversarial Examples to Evade Statistical Detectors

[T4] **Amazon Web Services Themis Team** (Guest Speaker), November 2021
Detecting Adversarial Input Distributions via Layer-wise Statistics

[T3] **Amazon Web Services Science Tech Presentations** (Guest Speaker), July 2021
Wonders and Dangers of Input-Adaptive Neural Network Inference

[T2] **ICML** (Conference Presentation), July 2021
When Does Data Augmentation Help With Membership Inference Attacks?

[T1] **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