

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

[F-5] RedactBench: A Formal Framework For LLM-based Confidential Information Redaction 2025
US IC Postdoctoral Fellowship — Third-Year Extension. Authored proposal after discussions with US Government sponsors; awarded ~\$100,000.

[F-4] Combating False Positives in ML-Based Security Applications With Context-Adaptive Classification 2024
Amazon Research Awards. Led proposal writing; secured \$80,000 (and \$20,000 in cloud credits) for the lab.

[F-3] Adaptable Machine Learning Systems for Antifragile Cyber Defenses 2023
US IC Postdoctoral Fellowship. Built on [P-12], [P-10] and [P-8]. Fellowship fully funded postdoc at UCSB for two years; awarded ~\$200,000.

[F-2] Distinct Impact of Trojans on the Internal Behaviors of Deep Neural Networks 2019
IARPA, Trojans in Artificial Intelligence (TrojAI). Based on findings in [P-2]. Contributed to lab funding and competition participation.

[F-1] Functional and Semantic Understanding of Deep Learning 2018
Laboratory for Telecommunication Sciences (LTS). Continuation of the research line in [P-2]; funded Ph.D. research.

Selected Publications

- [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; **To appear in IEEE S&P 2026**
- [P-12] “PoisonedParrot: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models”
Michael-Andrei Panaiteescu-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 Panaiteescu-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, Krishnamurthy Kenthapadi; **ICML 2022**
- [P-7] “Qu-ANTI-zation: Exploiting Quantization Artifacts for Achieving Adversarial Outcomes”
Sanghyun Hong, Michael-Andrei Panaiteescu-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**

Under Submission Pre-Prints

- [U-3] “Characterizing and Detecting Harassment in Social Virtual Reality”
Zihao Su, Kyle Zeng, Dongyu Meng, **Yigitcan Kaya**, Gianluca Stringhini, Christopher Kruegel, Giovanni Vigna
- [U-2] “RFC-Agent: An RFC-Aware Multi-Agent Reasoning System for Network Protocol Security Analysis”
Stijn Pletinckx, **Yigitcan Kaya**, Wenbo Guo, Christopher Kruegel, Giovanni Vigna
- [U-1] “From Documentation to Zero-day Vulnerabilities: LLM-Driven Fuzzing of Javascript Engines in PDF Readers”
Suyue Guo, Stijn Pletinckx, Tianle Yu, **Yigitcan Kaya**, Wenbo Guo, Christopher Kruegel, Giovanni Vigna

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) Advised 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) Advised seven research interns under the NSF ACTION Institute on three projects involving security and privacy issues of large language models; one of the projects formed the basis of [F-5] and the other led to a NeurIPS'25 workshop publication.
- (2024) Advised five research interns under the NSF ACTION Institute on security vulnerabilities of AI chatbots on the web; project led to publication [P-13].
- (2021) Advised 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-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

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