Yigitcan Kaya
Postdoctoral Fellow – University of California, Santa Barbara

Research Statement

AI systems have moved from research prototypes to core components of modern infrastructure. They now drive malware detection, customer support, content moderation, and program analysis. Yet the way we evaluate these systems has not kept pace. Most assessments still rely on clean inputs, fixed datasets, and stable operating conditions. These assumptions rarely hold in real deployments. When confronted with adversaries, shifting data distributions, or conflicting operational constraints, AI components often fail in ways that benchmarks do not reveal, precisely when reliability matters most.

My research lies at the intersection of **AI and systems security**, where I study how AI-driven components behave under the *actual* conditions found in security-critical environments. I focus on a central problem: **the persistent gap between how we evaluate AI models and how they fail in the wild**. Using empirical, measurement-driven methods, I build AI systems that are *robust* (resilient to adversarial or unexpected inputs) and *adaptable* (able to improve using deployment-time feedback). This *measurement-first philosophy* guides my work: before proposing defenses, I uncover how and why systems fail in practice. Following this approach, my research advances along three pillars: ① rigorous, adversary-aware evaluation, ② empirical analysis of deployment-time failures; and ③ practical defenses grounded in these measurements.

This strategy has uncovered several previously unknown failure modes: fairness interventions that inadvertently increase adversarial vulnerability [1]; malware detectors whose real-world accuracy drops by more than 70 percentage points [2]; and systemic weaknesses in widely deployed AI applications [3]. By combining expertise in AI, systems, and security, I have introduced new vulnerability classes (*overthinking* [4], *inconspicuous data poisoning* [8]), developed practical defenses [5, 9], and mapped trade-offs between competing defensive goals [6]. These findings have influenced follow-up work in both AI and security communities [18, 19, 20].

Across my projects, I approach AI with a security-informed, skeptical lens. A recurring theme in my work is that *old threats resurface in new forms* once AI becomes part of the systems stack. For example, I was among the first to document AI analogues of denial-of-service attacks [6], mimicry-style evasion [7], and client-side state manipulation vulnerabilities in live chatbot systems [3]. This perspective reflects training in world-class security research groups throughout my PhD and postdoc, as well as collaborations with industry teams whose insights have directly inspired multiple projects [2, 14].

My research has appeared in leading AI and security venues, including IEEE S&P, ICML, ICLR, USENIX Security, NeurIPS, SaTML, and NAACL. I have mentored junior researchers across several of these projects, helping them develop their own research trajectories. My work has also attracted substantial external support: my PhD was fully funded through two proposals based on my publications, and my postdoc is supported by the competitive U.S. Intelligence Community Postdoctoral Fellowship. I additionally led a successful Amazon Research Award proposal that initiated an ongoing collaboration [14]. Beyond academia, my work has been featured in *MIT Technology Review* [21], *VentureBeat* [22], and other media outlets. During two AWS internships, I contributed to patent applications for planned AI-security products.

In what follows, I describe my research across the three pillars: ① adversarial evaluation, ② deployment-time failure analysis, and ③ the translation of these insights into practical defenses. I conclude with my long-term vision: building evaluation frameworks and adaptive systems that close the research–practice gap and safeguard AI components in critical applications such as cyber defense and software development.

① Rigorous, Adversary-Aware Evaluation of AI Systems

Before we can credibly defend AI systems, we must understand how adversaries can break them. Early work in adversarial machine learning established influential threat models such as adversarial examples [17], but these capture only a narrow subset of real-world failure modes. *My research expands this landscape by applying security principles, such as careful attack surface analysis, to AI evaluation.* Through this perspective, I have uncovered overlooked vulnerabilities, developed practical attack methods to measure defensive progress, and revealed how classical security threats re-emerge in AI pipelines. Together, these studies have demonstrated that comprehensive threat modeling is essential for real-world security.

Slowdown attacks on input-adaptive inference [4, 6, 10]. Availability is a fundamental security property that has been largely absent from AI evaluations. As low-latency inference becomes critical, particularly on edge devices, adversaries gain incentives to exploit latency-sensitive components. My work provides the first systematic study of *slowdown attacks* against input-adaptive inference, a popular technique that reduces unnecessary computation by mitigating *overthinking* (an AI pathology I defined in [4]). I showed that carefully crafted inputs can force models to execute maximal computation, increasing latency by up to 5× in edge deployments. More importantly, I discovered a structural tension: adversarial training (the standard defense) offers limited protection, while slowdown-specific defenses fail against classical adversarial examples. This conflict between robustness objectives motivated follow-up work that investigated defenses [20] and uncovered additional weaknesses using similar methodology [19].

Mimicry attacks against adversarial example detectors [7]. Detection-based defenses are widely studied for countering adversarial attacks, yet most can be evaded by hand-crafted *adaptive* attacks [23]. During an AWS internship, I developed the *Statistical Indistinguishability Attack* (SIA), a general-purpose adaptive attack designed as a rigorous benchmark for adversarial-example detectors. Inspired by classical mimicry attacks, SIA generates adversarial inputs whose statistical properties match those of benign data, thereby evading detectors by construction. I demonstrated that SIA compromises a broad set of published detectors (e.g., [24, 25]) without detector-specific tuning, while revealing fundamental trade-offs between detectability and attack success. These insights clarified the limits of detection-based defenses and generated significant interest within AWS, culminating in a practitioner-focused article on the AWS AI Blog [26].

Fairness methods as attack amplifiers [1]. Group-robustness methods [27] aim to improve fairness by up-weighting underrepresented groups during training. My work exposed a critical blind spot: these methods cannot distinguish genuine minority examples from adversarial poisons in the training set. As a result, they can inadvertently amplify poisoning attacks, boosting attacker success rates from 0% to 97%. Conversely, poisoning defenses that remove suspected attacks also discard minority samples, increasing performance disparity by nearly 50%. These findings show that fairness and security goals can directly conflict, and that benchmark-centric evaluations obscure such conflicts.

Expanding threat models [11, 12, 8, 13]. I have expanded AI threat modeling along two dimensions central to my research vision. First, I helped demonstrate that AI systems are uniquely susceptible to classical hardware attacks. Because models primarily store parameters rather than code, blind Rowhammer attacks can silently flip bits in memory and silently destroy AI accuracy without crashing the system [11]. Repeated inference operations also leak timing patterns, enabling cache-side-channel recovery of proprietary architectures [12]. Second, I contributed to developing realistic data poisoning threats targeting modern AI systems trained on unverified web data. This work introduced new attack classes: inconspicuous poisoning that evades detection [8] and attacks that induce copyright violations in LLMs [13]. Both lines of research show how applying security insights to AI reveals emerging vulnerabilities invisible to conventional threat models.

Impact. Several methods from this research direction have become prototypical tools for subsequent work. The Shallow-Deep Network architecture I designed to study overthinking [4] has been widely adopted in AI, systems, and security research, enabling investigations into overthinking in language models [18], efficient model architectures [28], and fault-tolerant AI systems [29]. Similarly, slowdown attacks [6], Rowhammer-based attacks [11], and inconspicuous poisoning [8] have collectively accumulated over 800 citations and now serve as standard threat models in AI security.

② When AI Systems Break: Failures in Deployment

As AI systems enter security-critical domains, a persistent gap emerges between laboratory performance and real-world reliability. Benchmarks capture known challenges, but deployments expose a *long tail* of unforeseen conditions, including distribution shifts, user-driven errors, and adversarial adaptations. *My research examines how AI systems fail in the wild, revealing vulnerabilities that controlled evaluations systematically miss.* These studies shift the focus from reactive patching after failures occur to proactively anticipating them before systems are deployed.

The sandbox-to-endpoint gap in malware detection [2]. Malware detection is one of AI's highest-stakes applications. Yet a persistent disconnect remains: detectors perform nearly flawlessly in academic studies, while practitioners report frequent failures in production. Through collaboration with industry partners, I analyzed this gap for behavioral malware detectors, a widely used class of systems. I found that researchers train and evaluate detectors on program behaviors collected in sandboxes (controlled, instrumented environments). In contrast, deployed systems must classify behaviors from real user machines, which follow an entirely different distribution. This distribution shift is severe: detectors with 95% recall on sandbox data drop to 20% on endpoint data. Even retraining on endpoint data raises recall only to 50%, showing that the real-world task is fundamentally harder than benchmarks suggest. Through careful measurement and new ML techniques, we improved performance by 5–30% over baselines, but the broader lesson is clear: behavior-based malware detection remains an open, unsolved problem despite optimistic laboratory claims.

Prompt injection vulnerabilities in deployed chatbots [3]. Prompt injection enables adversaries to manipulate AI agents through malicious instructions. To mitigate this threat, developers train models with *instruction hierarchies* [30] that prioritize system-level prompts over user inputs. My work was the first to evaluate these defenses in real-world web deployments, focusing on third-party chatbot plugins used on more than 10,000 websites across government, education, and e-commerce domains. I discovered that these plugins inherit a classical web vulnerability: client-side state manipulation. By modifying local message histories, attackers can inject privileged instructions, bypassing developer controls and increasing attack success by up to 5×. Prior research implicitly assumed secure implementations; I showed that application-level flaws invalidate that assumption, leading to a systematic underestimation of risk. This study exemplifies my measurement-first approach: examining live deployments reveals failure modes invisible to controlled evaluations.

Impact. To foster progress, I released the *Malware-in-the-Wild Benchmark*, the first testbed designed to simulate endpoint conditions and bridge the research–practice divide. Our dataset has already been shared with five research groups, moving the community toward more realistic evaluations. For the chatbot study, our responsible disclosures prompted four major plugin vendors to issue critical security patches, improving the security of thousands of deployed chatbots on the web.

③ Building Adaptable AI Systems

Most prior research prepares AI systems for *anticipated* threats, for example, through adversarial training or data augmentation. However, my work shows that these methods introduce hidden trade-offs, and real deployments expose a long tail of unpredictable failures. *My research translates these empirical insights into a new design principle: building adaptable AI systems that leverage deployment-time feedback to improve their robustness and reliability.*

Addressing policy discrepancy in cyber defenses [14]. AI-based cyber defenses are typically trained on historical samples labeled as malicious or benign. However, discussions with practitioners revealed a critical gap: labels are often context-dependent rather than inherent properties. For example, analysts frequently mark legitimate alerts as “false positives” for business or workflow reasons, and prior evidence shows that most false positives arise this way [31]. My research formalizes this issue as *policy discrepancy* and develops practical methods to mitigate it. We train models that quickly adapt to an organization’s active policy using only a small set of labeled samples collected during deployment. Our approach recovers roughly 50% of the accuracy lost to policy discrepancy with as few as 25 samples. This enables defenses that avoid predicting inherently ambiguous conditions in advance and instead adapt to them on the fly.

LLM agents for security testing [15, 16]. LLMs offer promise for automating security analysis, but their static predictions often hallucinate, overlook edge cases, or misinterpret specifications. During my postdoc, I worked with junior PhD students to build LLM-based agents that systematically test PDF readers [15] and network protocol implementations [16]. Our key idea is to transform LLMs from static classifiers into *adaptive testers*: agents that iteratively refine hypotheses, verify results, retrieve grounding evidence, and learn from failures through feedback loops. This adaptive structure enables deeper, more reliable vulnerability discovery than one-shot prompting.

Impact. To advance post-deployment adaptation for AI-based security systems, I led a successful Amazon Research Award that now supports an ongoing collaboration. Our adaptive testing systems have already uncovered multiple vulnerabilities, including critical zero-days, in widely used software, prompting swift vendor patches through responsible disclosure.

Future Research Directions

AI is reshaping computer security, offering powerful capabilities while introducing new classes of risk. My work has revealed systematic gaps in how we evaluate, deploy, and safeguard these systems. Building on these insights, my long-term goal is to establish a research group focused on *adaptive and antifragile* AI: systems that not only withstand uncertainty but improve when exposed to it. I will pursue this vision along two complementary directions.

Near-Term Directions: Designing AI Systems for End-to-End Security

AI components rarely operate in isolation. In deployed settings, malware detectors, code analyzers, and chatbot moderators interact with rule-based filters, heuristics, and human analysts. Yet most research evaluates AI modules independently, obscuring the interactions that shape real-world security outcomes. My near-term work aims to close this gap by *embedding AI within full defense pipelines* and addressing the data limitations that constrain realistic evaluation.

Optimizing end-to-end systems. Real deployments must balance multiple objectives, including accuracy, latency, scalability, transparency, and cost. I will extend the *learning with privileged information* paradigm [32] to train models using rich features available only during training (e.g., detailed program traces) while remaining lightweight at deployment. Beyond improving individual components, I aim to co-train AI and non-AI modules to decide when to delegate decisions, calibrate confidence, and optimize system-level trade-offs. This reframes robustness as a property of the *entire defense pipeline*, not of any single model.

Bridging the data gap. The industry possesses realistic, large-scale data, while academia has the capacity to innovate. My sandbox-to-endpoint study [2] highlighted this divide: real endpoints revealed failures invisible in academic benchmarks, yet the underlying data could not be publicly shared. To reconcile this, I will develop *synthetic dataset generators* that reproduce the statistical and operational properties of proprietary environments while preserving confidentiality. This will enable safe data sharing, reproducible evaluation, and more meaningful collaboration between academia and industry.

Long-Term Vision: Toward Antifragile AI Agents

As AI systems take on open-ended tasks, such as security triage, code generation, and autonomous decision-making, they increasingly operate under underspecified, shifting, or conflicting human goals. Robustness alone is insufficient; *future AI systems must be antifragile: capable of learning from ambiguity and improving through deployment feedback*. My work on policy discrepancy [14] provides the foundation, showing that models can adapt to new organizational contexts that were never fully defined during training. The next challenge is to generalize this principle toward AI systems that recognize uncertainty, seek clarification, and refine their own assumptions.

AI-assisted development as a testbed. Modern software development is rapidly evolving into a workflow in which AI systems translate human intent directly into code. Security risks now stem as much from *semantic ambiguity* as from implementation errors. For example, if an e-commerce system does not specify whether coupons are single-use or reusable, an AI agent may infer an unintended policy from training data, creating vulnerabilities. Most AI systems optimize for seamless interaction rather than clarification, thereby amplifying these risks. Like policy discrepancy, this setting highlights a frontier where failures arise from ambiguous goals, not faulty algorithms.

A unified framework for operating under uncertainty. I aim to develop a framework for antifragile AI that: (1) measures how systems form assumptions under underspecification, (2) surfaces those assumptions for human review, and (3) verifies alignment during deployment. Through iterative clarification and feedback, AI systems can learn which ambiguities matter and improve over time. This vision spans domains where AI must act with incomplete information, including code generation, autonomous systems, and cyber defense. Ultimately, I seek to understand AI not only as a target for vulnerabilities but as a source of failure modes, and to build systems that are secure, adaptive, and self-correcting.

Collaboration and Funding Plan

I will drive my research agenda forward through a strong network across academia, industry, and government. I collaborate with former mentees now at leading universities; researchers in my postdoc group SecLab, a world-class systems security lab that leads the NSF ACTION Institute and whose alumni hold faculty positions across top U.S. intitutions; and senior faculty whose work has shaped the AI–security landscape, including David Wagner (UC Berkeley) and Lorenzo Cavallaro (UCL). I maintain ties with my PhD group at the University of Maryland and partner with industry through an Amazon Research Award that supports our work on post-deployment adaptation. I also work with researchers in the U.S. Intelligence Community, where we have identified high-impact problems at the intersection of AI security and national priorities. These efforts have led to a third-year extension of my postdoctoral fellowship to advance LLM-based information-redaction systems. Together, these relationships provide a strong foundation to accelerate my group's research and amplify its impact.

In the first three years, I will target early-career programs in security and national defense open to U.S. citizens, including the DoD Young Investigator Programs (AFOSR, ARO, ONR) and the DARPA Young Faculty Award, emphasizing my expertise in AI-driven cyber defense. In parallel, I will leverage existing collaborations with senior faculty to co-lead proposals to NSF Secure and Trustworthy Cyberspace (SaTC) and related programs, as well as industry research awards in AI security (e.g., from Amazon or Google). By year five, I expect to integrate these efforts into a competitive NSF CAREER proposal centered on my long-term vision for adaptive and antifragile AI in security-critical systems.

Conference Publications

[1] Michael-Andrei Panaitescu-Liess*, **Yigitcan Kaya***, Sicheng Zhu, Furong Huang, and Tudor Dumitras. "Like Oil and Water: Group Robustness Methods and Poisoning Defenses May Be at Odds". In: *2024 International Conference on Learning Representations (ICLR)*.

[2] **Yigitcan Kaya**, Yizheng Chen, Marcus Botacin, Shoumik Saha, Fabio Pierazzi, Lorenzo Cavallaro, David Wagner, and Tudor Dumitras. "ML-Based Behavioral Malware Detection Is Far From a Solved Problem". In: *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*.

[3] **Yigitcan Kaya**, Anton Landerer, Stijn Pletinckx, Michelle Zimmermann, Christopher Kruegel, and Giovanni Vigna. "When AI Meets the Web: Prompt Injection Risks in Third-Party AI Chatbot Plugins". In: *2026 IEEE Symposium on Security and Privacy (S&P)*.

[4] **Yigitcan Kaya**, Sanghyun Hong, and Tudor Dumitras. "Shallow-Deep Networks: Understanding and Mitigating Network Overthinking". In: *2019 International Conference on Machine Learning (ICML)*.

[5] **Yigitcan Kaya** and Tudor Dumitras. "When Does Data Augmentation Help With Membership Inference Attacks?" In: *2021 International Conference on Machine Learning (ICML)*.

[6] Sanghyun Hong*, **Yigitcan Kaya***, Ionuț-Vlad Modoranu, and Tudor Dumitras. "A Panda? No, It's a Sloth: Slowdown Attacks on Adaptive Multi-Exit Neural Network Inference". In: *2021 International Conference on Learning Representations (ICLR)*.

[7] **Yigitcan Kaya**, Muhammad Bilal Zafar, Sergul Aydore, Nathalie Rauschmayr, and Krishnaram Kenthapadi. "Generating Distributional Adversarial Examples to Evade Statistical Detectors". In: *2022 International Conference on Machine Learning (ICML)*.

[8] Octavian Suciu, Radu Marginean, **Yigitcan Kaya**, Hal Daume III, and Tudor Dumitras. "When Does Machine Learning FAIL? Generalized Transferability for Evasion and Poisoning Attacks". In: *2018 USENIX Security Symposium*.

[9] Shoumik Saha, Wenxiao Wang, **Yigitcan Kaya**, Soheil Feizi, and Tudor Dumitras. "DRSM: De-Randomized Smoothing on Malware Classifier Providing Certified Robustness". In: *2024 International Conference on Learning Representations (ICLR)*.

[10] Kamala Varma, Arda Numanoglu, **Yigitcan Kaya**, and Tudor Dumitras. "Understanding, Uncovering, and Mitigating the Causes of Inference Slowdown for Language Models". In: *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*.

[11] Sanghyun Hong, Pietro Frigo, **Yigitcan Kaya**, Cristiano Giuffrida, and Tudor Dumitraș. "Terminal Brain Damage: Exposing the Graceless Degradation in Deep Neural Networks Under Hardware Fault Attacks". In: *2019 USENIX Security Symposium*.

[12] Sanghyun Hong, Michael Davinroy, **Yigitcan Kaya**, Dana Dachman-Soled, and Tudor Dumitras. "How to 0wn the NAS in Your Spare Time". In: *2020 International Conference on Learning Representations (ICLR)*.

[13] Michael-Andrei Panaitescu-Liess, Pankayaraj Pathmanathan, **Yigitcan Kaya**, Zora Che, Bang An, Sicheng Zhu, Aakriti Agrawal, and Furong Huang. "PoisonedParrot: Subtle Data Poisoning Attacks to Elicit Copyright-Infringing Content from Large Language Models". In: *2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*.

Preprints

[14] **Yigitcan Kaya**, Baris Coskun, Wenbo Guo, Christopher Kruegel, and Giovanni Vigna. *One Sample, Many Truths: Using Context-Adaptive Classification to Combat Security Policy Discrepancy*. Preprint (Under review) – Received Amazon Research Award in September 2024.

[15] Suyue Guo, Stijn Pletinckx, Tianle Yu, **Yigitcan Kaya**, Wenbo Guo, Christopher Kruegel, and Giovanni Vigna. *From Documentation to Zero-day Vulnerabilities: LLM-Driven Fuzzing of Javascript Engines in PDF Readers*. Preprint (Under review).

[16] Stijn Pletinckx, **Yigitcan Kaya**, Wenbo Guo, Christopher Kruegel, and Giovanni Vigna. *RFC-Agent: An RFC-Aware Multi-Agent Reasoning System for Network Protocol Security Analysis*. Preprint (Under review).

Additional References

[17] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. “Intriguing Properties of Neural Networks”. In: *2014 International Conference on Learning Representations (ICLR)*.

[18] Wangchunshu Zhou, Canwen Xu, Tao Ge, Julian McAuley, Ke Xu, and Furu Wei. “BERT Loses Patience: Fast and Robust Inference with Early Exit”. In: *2020 Advances in Neural Information Processing Systems (NeurIPS)*.

[19] Zheng Li, Yiyong Liu, Xinlei He, Ning Yu, Michael Backes, and Yang Zhang. “Auditing Membership Leakages of Multi-Exit Networks”. In: *2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*.

[20] Ravishka Rathnasuriya, Tingxi Li, Zexin Xu, Zihe Song, Mirazul Haque, Simin Chen, and Wei Yang. “SoK: Efficiency Robustness of Dynamic Deep Learning Systems”. In: *2025 USENIX Security Symposium*.

[21] Karen Hao. *AI consumes a lot of energy. Hackers could make it consume more.* May 2021. URL: https://www.technologyreview.com/2021/05/06/1024654/ai-energy-hack-adversarial-attack/.

[22] Ben Dickson. *Machine Learning Security Needs New Perspectives and Incentives*. June 2021. URL: https://venturebeat.com/2021/06/06/machine-learning-security-needs-new-perspectives-and-incentives/.

[23] Nicholas Carlini and David Wagner. “Adversarial Examples Are Not Easily Detected: Bypassing Ten Detection Methods”. In: *2017 ACM Workshop on Artificial Intelligence and Security (AISec)*.

[24] Jayaram Raghuram, Varun Chandrasekaran, Somesh Jha, and Suman Banerjee. “A General Framework for Detecting Anomalous Inputs to DNN Classifiers”. In: *2021 International Conference on Machine Learning (ICML)*.

[25] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. “The Odds Are Odd: A Statistical Test for Detecting Adversarial Examples”. In: *2019 International Conference on Machine Learning (ICML)*.

[26] Nathalie Rauschmayr, Sergul Aydore, **Yigitcan Kaya**, and Bilal Zafar. *Detect adversarial inputs using Amazon SageMaker Model Monitor and Amazon SageMaker Debugger*. Amazon Web Services. Dec. 2020. URL: https://aws.amazon.com/blogs/machine-learning/detect-adversarial-inputs-using-amazon-sagemaker-model-monitor-and-amazon-sagemaker-debugger/.

[27] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. “Fairness Without Demographics in Repeated Loss Minimization”. In: *2018 International Conference on Machine Learning (ICML)*.

[30] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. “The Instruction Hierarchy: Training LLMs to Prioritize Privileged Instructions”. In: *arXiv preprint arXiv:2404.13208* (2024).

[31] Bushra A Alahmadi, Louise Axon, and Ivan Martinovic. “99% False Positives: A Qualitative Study of SOC Analysts’ Perspectives on Security Alarms”. In: *2022 USENIX Security Symposium*.

[32] Vladimir Vapnik and Akshay Vashist. “A New Learning Paradigm: Learning Using Privileged Information”. In: *Neural networks* 22.5-6 (2009), pp. 544–557.