Yigitcan Kaya
Postdoctoral Fellow – University of California, Santa Barbara

Research Statement

AI systems fail in the wild in ways that benchmarks never predict. We are witnessing a rapid transition from research prototypes to critical infrastructure. AI now powers malware detection, customer service, content moderation, and program analysis across thousands of sensitive applications. Yet evaluation remains rooted in standardized benchmarks that assume clean data, isolated metrics, and stable environments. This disconnect between laboratory conditions and reality leads to dangerous overestimation of robustness at a time when the stakes have never been higher.

My research directly confronts this gap. I build AI that is both *robust* (withstanding unexpected conditions) and *adaptable* (improving through deployment feedback), whether facing ambiguity, distribution shifts, adversarial pressure, or conflicting objectives. I pursue this through a *measurement-first* philosophy: before proposing defenses, I identify how systems actually break. My research proceeds through three pillars: **(1)** developing rigorous evaluations under adversarial pressure, **(2)** investigating how deployed systems fail under unforeseen conditions, and **(3)** translating these insights into practical defenses.

This approach has revealed fundamental limitations invisible in standard benchmarks: defenses that amplify the attacks they are meant to stop [1], detection systems with 70-point accuracy drops at deployment [2], and emergent vulnerabilities in widely-deployed applications [3]. Combining AI expertise with hands-on experience at leading security labs, I have coined new classes of vulnerabilities (overthinking [4], inconspicuous poisoning [8]), designed practical defenses for deployed systems [5, 9], and exposed critical trade-offs between competing defensive goals [6]. This work has sparked follow-up research by others addressing these tensions [18, 20] and uncovering new ones through similar measurement-first analyses [19].

I study AI through a skeptical lens informed by security domain knowledge. A recurring theme in my research is demonstrating how *old becomes new*: how classical security threats re-emerge in different forms against AI systems. For example, I was among the first to study analogues of denial-of-service [6] and mimicry [7] attacks in the AI context, and to show how thousands of chatbots remain vulnerable to client-side state manipulation [3]. This perspective stems from extensive experience in world-class security research environments throughout my PhD and postdoc. I also actively seek opportunities to engage with practitioners, whose real-world insights have inspired multiple projects [2, 14].

My research has resulted in first-authored publications at top venues spanning AI and security, including ICML, ICLR, IEEE S&P, and SaTML, with additional contributions at NeurIPS, USENIX Security, and NAACL. In several of these projects, I mentored junior PhD students and research interns as they developed their own research vision. My work has attracted significant external support: my PhD was fully funded through two proposals based on my publications, and my postdoc is entirely self-funded through a competitive U.S. Government fellowship (~$285,000 over three years). I also led a successful Amazon Research Award proposal ($80,000) that launched an ongoing collaboration [14]. Beyond academia, my work has been featured in *MIT Technology Review* [21], *VentureBeat* [22], and other media outlets. During my AWS internships, it produced two patent applications and the seeds of a new AI security offering.

In the following, I describe my research across three pillars: adversarial evaluation, studies of deployment failures, and the translation of insights into practical defenses. I then outline my broader vision: developing evaluation frameworks that expose vulnerabilities, bridge research-practice gaps, and ultimately safeguard AI systems in critical applications such as cyber defense and software development.

Evaluating AI Systems Under Adversarial Pressure

Before defending AI systems, we must understand how adversaries may break them. Early work in adversarial machine learning established threat models like *adversarial examples* [17], but these capture only a narrow slice of real-world threats. My research expands this landscape by applying security domain knowledge to expose vulnerabilities that standard evaluations miss. This has uncovered new vulnerability classes, enabled practical attacks for measuring defensive progress, and revealed how classical security threats resurface in AI systems. Below, I discuss my strongest contributions in this direction, which demonstrate that comprehensive threat modeling is essential for real-world security.

Slowdown attacks on input-adaptive inference [6, 10]. Adversaries routinely target system availability through denial-of-service attacks. As AI systems increasingly prioritize low latency, especially at the edge, this threat becomes critical. My work was among the first to systematically study this risk against input-adaptive inference, a prominent approach for reducing latency. Such techniques prevent *overthinking* (an AI pathology I defined [4]) by enabling fast predictions on simple inputs while maintaining accuracy on complex ones. I showed that adversaries can craft inputs that force maximum computation, increasing latency by up to 5× in edge deployments. More importantly, I discovered a fundamental tension: adversarial training (the standard defense) provides minimal protection against slowdown attacks, while slowdown-specific defenses fail against conventional attacks. This exposed an inherent trade-off between robustness objectives and inspired follow-up work by others that investigated defenses [20], and identified new vulnerabilities using a similar methodology [19].

Mimicry attacks against detectors [7]. Detection-based defenses are widely studied for countering adversarial attacks, yet most can be evaded by hand-crafted *adaptive* attacks [23]. My work, conducted during an AWS internship, developed the Statistical Indistinguishability Attack (SIA), a generic adaptive attack serving as a rigorous benchmark for detectors. Inspired by classical mimicry attacks, SIA constrains adversarial samples to follow the distribution of clean inputs, enabling an intuitive trade-off between evasiveness and effectiveness. I demonstrated SIA's power by breaking a wide range of published detectors (e.g., [24, 25]) without customization, while identifying detector properties that force attacks to compromise between success and detectability. These insights highlight promising directions for building detectors that resist statistical mimicry. This work sparked interest in AI security at AWS, leading to a practitioner-focused post on the AWS AI Blog [26].

Fairness methods as attack amplifiers [1]. Group robustness methods aim to make AI models fairer by reducing error disparities across demographic groups through the amplification of underrepresented training samples. My work exposed a fundamental tension: these methods cannot distinguish legitimate minority samples from adversarial poisons in the training set, amplifying attacks and boosting adversary success from 0% to 97%. Conversely, poisoning defenses that remove suspected attacks also eliminate minority samples, amplifying error discrepancy by nearly 50%. These methods pursue opposing goals with conflicting heuristics, and combining them fails to resolve the trade-off. This work exemplifies how benchmark-driven evaluations obscure critical trade-offs with real-world consequences.

New threat models [11, 12, 8, 13]. I have advanced threat modeling in two directions aligned with my research vision. First, I contributed to work showing that AI systems are under amplified risk from classical hardware attacks, namely Rowhammer attacks (which exploit hardware to flip bits in memory) and cache-side channel attacks (which leak information through timing patterns in shared CPU caches). Unlike traditional software that stores instructions, AI primarily stores parameters in memory, enabling blind Rowhammer attacks to cause catastrophic accuracy loss without crashing the program [11]. Similarly, AI systems repeat identical computations during inference, allowing attackers to steal proprietary model architectures via cache

side-channels [12]. Second, I contributed to developing realistic poisoning attacks against modern AI systems trained on unverified web data. These introduced new threat models: inconspicuous poisoning that evades detection [8] and attacks that induce copyright infringement in LLMs [13]. Both lines exemplify applying security insights to expose emerging vulnerabilities invisible to conventional threat models.

When AI Systems Break: Deployment Failures

As AI systems are increasingly deployed in security-critical settings, a dangerous gap emerges between laboratory performance and real-world reliability. Standard benchmarks optimize for known challenges, but deployment introduces a *long tail* of unforeseen problems, including distribution shifts, user errors, and adversarial adaptations. My research investigates how deployed AI systems actually fail in the wild, revealing vulnerabilities that controlled evaluations systematically miss. These insights enable a shift from reactive patches to proactive defenses for catching failures before deployment, not after.

The sandbox-to-endpoint gap in malware detection [2]. Malware detection is one of AI's highest-stakes applications. Yet a troubling disconnect exists: detectors achieve near-perfect accuracy in prior evaluations while practitioners report frequent failures in the field. Building on industry partnerships, my research demystified this gap for behavioral malware detectors, a widely deployed technology. I discovered that research trains and evaluates detectors on program behaviors from sandboxes (controlled environments), but deployed detectors must analyze behaviors from actual user machines, which follows an entirely different distribution. This shift is catastrophic: detectors with 95% recall on sandbox data drop to 20% on endpoint data. Even retraining detectors directly on endpoint data achieves only 50% recall, revealing that the real-world task is fundamentally harder than what research benchmarks suggest. Through careful measurement and new ML techniques, we improved performance by 5–30% over baselines, but the core lesson is clear: behavior-based malware detection remains unsolved despite claims in prior work. To enable progress, I released a benchmark simulating real endpoint data that allows testing under production conditions, bridging the research-practice gap that standard evaluations ignore.

Prompt injection vulnerabilities in deployed chatbots [3]. Prompt injection allows adversaries to manipulate AI agents into executing malicious instructions. To defend against this, AI companies train models with *instruction hierarchy*, prioritizing developer instructions over user inputs. While advanced AI applications receive significant security attention, my work was the first to examine real-world AI chatbots deployed via third-party web plugins, which power over 10,000 websites across critical domains such as government, education, and infrastructure. I discovered these chatbots are vulnerable to a classical web security flaw: client-side state manipulation, which allows adversaries to inject privileged instructions that bypass developer controls and built-in defenses, increasing attack success by up to 5×. Prior work assumed secure implementations, but I showed that application-level vulnerabilities grant adversaries unforeseen capabilities, causing a dangerous underestimation of real-world risk. This exemplifies my measurement-first philosophy: studying how systems break in the wild reveals gaps invisible to controlled evaluations. My responsible disclosure led to two major plugins issuing critical security fixes affecting thousands of websites.

Building Adaptable AI Systems

Most prior research builds AI systems that are robust to foreseeable threats through methods like adversarial training. However, my research has shown that existing robustness methods create invisible trade-offs against unforeseen threats, and deployment introduces a long tail of unpredictable failures. I translate these insights into practical defenses and a paradigm shift: from static robustness to *adaptable* AI systems that improve

under deployment conditions. This vision has been the primary focus of my postdoc, funded by a competitive US Government fellowship for my proposal.

Addressing policy shift in cyber defenses [14]. AI-based cyber defenses are trained on historical samples labeled as malicious or benign to recognize threats during deployment. However, a meeting with practitioners revealed a critical gap: labels are often context-dependent rather than inherent properties. For example, organizations may treat correct alerts as false positives for business reasons, and evidence suggests most false positives arise this way, creating discrepancy between measured and perceived performance. My research formally defines this as *policy shift* and develops a practical solution through adaptation. Our approach trains adaptable models that quickly tune to unknown organizational policies using small labeled samples from deployment. This approach allows us to avoid predicting inherently ambiguous deployment policies in advance and to build systems that improve by adapting to them. To further explore this direction, I led a successful Amazon Research Award proposal, enabling ongoing collaboration on policy shift challenges.

LLM agents for security testing [15, 16]. LLMs offer promise for automating security analysis but suffer from hallucinations that produce incomplete or false solutions. During my postdoc, I worked with junior PhD students to develop agentic systems that rigorously test PDF readers [15] and network protocols [16] for vulnerabilities. These systems discovered multiple zero-day vulnerabilities, leading to vendor patches. The key insight is to design systems that enable LLMs to continuously improve and adapt their outputs, even when starting from potentially faulty answers, by incorporating verification, retrieval-augmented generation, structured reasoning, and feedback loops from failures.

Future Research Directions

AI systems are transforming computer security, offering immense promise while introducing new hazards. Through my research, I have developed a vision for identifying gaps in how we safeguard these systems, anticipating emerging challenges, and building solutions that work in practice. I will build a research team pursuing this vision along the following themes.

Holistic Approach to Designing AI Systems in Security

AI systems in security never work in isolation; they operate within larger systems combining multiple AI and non-AI components. Yet research consistently isolates AI during evaluation, creating the gaps I have identified throughout my work. For example, real-world malware detection deploys rule-based and heuristic defenses as frontline filters, with AI complementing them, while most research evaluates single AI components tackling entire tasks. I attribute this disconnect to two critical gaps: lack of frameworks for developing AI within complete systems, and insufficient access to real-world datasets. I will address both challenges through the following research directions.

Optimizing end-to-end systems. Real deployments require optimizing multiple competing metrics, including latency, explainability, and cost, not just AI accuracy. I will explore the learning using privileged information paradigm [27] to train AI using expensive features (e.g., program behaviors) unavailable at deployment, enabling systems that balance performance across constraints. Beyond individual components, I will develop joint optimization techniques that co-train AI and non-AI components, learning where rule-based systems should hand off to AI and how to calibrate confidence thresholds across the entire pipeline. This approach shifts focus from maximizing single-component performance to maximizing end-to-end security posture.

Bridging the data gap. Industry holds real-world data but lacks research capacity; academia has researchers

but not the data. My sandbox-to-endpoint work [2] relied on proprietary industry data that exposed critical failures invisible in public benchmarks, yet could not be shared. I will develop synthetic data generation techniques that preserve statistical properties and deployment challenges while protecting proprietary information. This enables industry to share representative datasets without competitive risk, allowing researchers to develop solutions that work in practice and creating a mutually beneficial ecosystem for AI security research.

Antifragile AI Agents Operating Under Ambiguity

A fundamental challenge emerges as AI agents increasingly make decisions based on underspecified human requirements: they must infer intent from ambiguous instructions, introducing invisible assumptions that become sources of failure. Building robust AI that resists such failures is insufficient; we need systems that are *antifragile*, improving through exposure to ambiguity rather than merely tolerating it. My research vision is to build AI systems that operate reliably under inherent ambiguity, not by eliminating it, but by systematically surfacing and adapting to critical uncertainties.

My work on policy shift [14] demonstrated how AI-based cyber defenses can adapt to ambiguous organizational policies, where the same alert may be a true positive or false positive depending on business context. This established the foundation for building adaptable AI systems. I now aim to generalize this vision: building antifragile AI that not only adapts to ambiguity but actively improves through it, learning which uncertainties matter across diverse deployment contexts.

AI-assisted development as a critical testbed. Software development exemplifies this challenge at scale. As developers express requirements in natural language rather than code, security risks shift from classical vulnerabilities like memory-safety errors to logic flaws arising from misaligned assumptions between developer intent and AI implementation. Consider an e-commerce coupon system: are coupons single-use or reusable? If unspecified, AI infers behavior from training data, potentially introducing critical vulnerabilities. Concerningly, AI companies encourage such inferences to reduce friction, viewing clarification requests as harmful to user experience. Like policy shift, systems fail not from implementation errors, but from mismatches between implicit human intent and AI's inferred specifications.

A unified framework for operating under uncertainty. I will develop techniques that enable AI agents to operate antifragile across domains: (1) measuring how AI makes assumptions under underspecification, (2) designing methods to surface critical assumptions for human review, and (3) creating verification techniques that catch misalignment before deployment. Crucially, systems improve through iterative clarification, learning which ambiguities matter in practice rather than requiring exhaustive upfront specification. Applications span code generation, autonomous systems, content moderation, and security decision-making; any domain where AI must act on incomplete human intent. This extends my vision: understanding AI not just as a target of vulnerabilities, but as a component that introduces new failure modes requiring fundamentally different security paradigms.

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: *The Twelfth International Conference on Learning Representations (ICLR)*. 2024.

[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: *Proceedings of the 2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. 2025.

[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: *Proceedings of the 2026 IEEE Symposium on Security and Privacy (S&P)*. 2026.

[4] **Yigitcan Kaya**, Sanghyun Hong, and Tudor Dumitras. "Shallow-Deep Networks: Understanding and Mitigating Network Overthinking". In: *36th International Conference on Machine Learning (ICML)*. 2019.

[5] **Yigitcan Kaya** and Tudor Dumitras. "When Does Data Augmentation Help With Membership Inference Attacks?" In: *38th International Conference on Machine Learning (ICML)*. 2021.

[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: *9th International Conference on Learning Representations (ICLR)*. 2021.

[7] **Yigitcan Kaya**, Muhammad Bilal Zafar, Sergul Aydore, Nathalie Rauschmayr, and Krishnaram Kenthapadi. "Generating Distributional Adversarial Examples to Evade Statistical Detectors". In: *39th International Conference on Machine Learning (ICML)*. 2022.

[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: *27th USENIX Security Symposium*. 2018.

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

[10] Kamala Varma, Arda Numanoglu, **Yigitcan Kaya**, and Tudor Dumitras. "Understanding, uncovering, and mitigating the causes of inference slowdown for language models". In: *Proceedings of the 2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. 2024.

[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: *28th USENIX Security Symposium*. 2019.

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

[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: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*. 2025.

Preprints

[14] **Yigitcan Kaya**, Wenbo Guo, Baris Coskun, Christopher Kruegel, and Giovanni Vigna. *One Sample, Many Truths: Context-Adaptive Classification to Combat Policy Shifts in Security Tasks*. Work in progress (Received Amazon Research Award in 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. In: *The Second International Conference on Learning Representations (ICLR)*. 2014.

[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: *Advances in Neural Information Processing Systems*. 2020.

[19] Zheng Li, Yiyong Liu, Xinlei He, Ning Yu, Michael Backes, and Yang Zhang. "Auditing membership leakages of multi-exit networks". In: *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2022.

[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: *34th USENIX Security Symposium*. 2025.

[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: *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 2017.

[24] Jayaram Raghuram, Varun Chandrasekaran, Somesh Jha, and Suman Banerjee. "A general framework for detecting anomalous inputs to dnn classifiers". In: *38th International Conference on Machine Learning (ICML)*. 2021.

[25] Kevin Roth, Yannic Kilcher, and Thomas Hofmann. "The odds are odd: A statistical test for detecting adversarial examples". In: *36th International Conference on Machine Learning (ICML)*. 2019.

[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] Vladimir Vapnik and Akshay Vashist. "A new learning paradigm: Learning using privileged information". In: *Neural Networks* 22.5 (2009). Advances in Neural Networks Research, pp. 544–557.