Yigitcan Kaya
Postdoctoral Fellow – University of California, Santa Barbara

Research Statement

AI systems have evolved from research prototypes into critical infrastructure. They now drive malware detection, customer service, content moderation, and program analysis in high-stakes settings. Yet most our evaluations still assume clean data, isolated metrics, and stable environments; conditions far removed from reality. This mismatch creates dangerous overconfidence, causing systems to fail in deployment in ways benchmarks never predict, often when failure is most costly.

My research at the intersection of **AI and systems security** directly confronts this gap. I aim to build AI that is *robust* (withstanding unexpected conditions) and *adaptable* (improving via deployment feedback) under ambiguity, distribution shift, adversarial pressure, and competing objectives. I follow a *measurement-first* philosophy: before proposing defenses, I identify how systems actually break. This work proceeds along three pillars: **(1)** developing rigorous evaluations under adversarial pressure, **(2)** studying how deployed systems fail under unforeseen conditions, and **(3)** translating these insights into practical defenses.

This strategy has revealed blind spots benchmarks overlook: defenses that inadvertently amplify attacks [1], detection systems whose accuracy collapses by 70 points at deployment [2], and emergent vulnerabilities in widely deployed applications [3]. Leveraging AI expertise and hands-on security experience, I have introduced new vulnerability classes (overthinking [4], inconspicuous poisoning [8]), designed practical defenses [5, 9], and exposed trade-offs between competing defensive goals [6]. These findings have inspired others to address these tensions [18, 20] and uncover new ones through similar empirical 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 resurface in new 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 training 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 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. In several projects, I mentored junior PhD students and research interns as they developed their own research directions. 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. I also led a successful Amazon Research Award proposal 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, I contributed to two patent applications and to the seeds of a new AI security offering.

In what follows, I first describe my research across the 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 principles to AI evaluation. Through this lens, I have uncovered overlooked vulnerabilities, enabled practical attacks to measure defensive progress, and revealed how classical security threats re-emerge in AI systems. Together, these studies have demonstrated that comprehensive threat modeling is essential for real-world security.

Slowdown attacks on input-adaptive inference [6, 10]. Availability is a core security goal that is often overlooked in AI evaluation. Adversaries gain new incentives to target it as low-latency inference becomes critical, especially on edge devices. My work was the first to systematically study *slowdown attacks* against input-adaptive inference, a prominent latency-reduction technique that reduces *overthinking* (an AI pathology I defined [4]). I showed that adversarial inputs can force models to perform maximal computation, increasing latency by up to 5× in edge deployments. More importantly, I discovered a fundamental tension: adversarial training (the standard defense) offers minimal protection, while defenses tailored to slowdown attacks fail against conventional attacks. This finding exposed conflicting robustness objectives and inspired follow-ups that investigated defenses [20], and discovered new vulnerabilities through 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]. During an AWS internship, I developed the *Statistical Indistinguishability Attack* (SIA), a general, adaptive attack that serves as a rigorous benchmark for detectors. Inspired by mimicry attacks in classical security, SIA crafts adversarial inputs that match the statistical distribution of clean data. I demonstrated that SIA can break a wide range of published detectors (e.g., [24, 25]) without customization, while revealing detector properties that force a trade off between detectability and attack effectiveness. These findings pointed to directions toward resilient detectors and 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 [27] aim to equalize performance across demographic groups by up-weighting underrepresented samples. My work exposed a critical blind spot: these methods cannot distinguish genuine minority examples from adversarial poisons in the training set, inadvertently boosting attacker success from 0 to 97%. Conversely, poisoning defenses that remove suspected attacks also discard minority samples, increasing disparity by nearly 50%. These findings show that fairness and security goals are at odds, and that benchmark-centric evaluation obscures such conflicts.

Expanding threat models [11, 12, 8, 13]. I have advanced threat modeling in two directions 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 and destroy accuracy without crashing the program [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.

When AI Systems Break: Deployment Failures

As AI systems enter security-critical domains, a gap emerges between laboratory performance and real-world reliability. Benchmarks optimize for known challenges, but deployment introduces a *long tail* of unforeseen conditions, including distribution shifts, user errors, and adversarial adaptations. My research investigates how AI systems fail in the wild, uncovering vulnerabilities controlled evaluations overlook. These studies enable a proactive shift: from patching failures after deployment to anticipating them before deployment.

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 appear near-perfect in academic studies while practitioners report frequent failures in the field. Through collaboration with industry partners, I analyzed

this gap for behavioral malware detectors, a widely deployed technology. I found that researchers train and evaluate detectors on program behaviors observed in sandboxes (controlled environments), but deployed systems must classify behaviors from real user machines, which follow an entirely different distribution. This shift is catastrophic: detectors with 95% recall on sandbox data drop to 20% on endpoint data. Even retraining detectors on endpoint data achieves only 50% recall, showing that the real-world task is fundamentally harder than what prior 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 optimistic claims. To foster progress, I released the *Malware-in-the-Wild Benchmark*, which simulates endpoint conditions and provides the first testbed bridging the research–practice divide.

Prompt injection vulnerabilities in deployed chatbots [3]. Prompt injection allows adversaries to manipulate AI agents through malicious instructions. To mitigate this, developers train models with *instruction hierarchies* [28] that prioritize developer prompts over user inputs. My work was the first to test these defenses in real-world web deployments, focusing on third-party chatbot plugins integrated into more than 10,000 websites across government, education, and e-commerce domains. I discovered these plugins inherit classical web vulnerability: client-side state manipulation. By modifying local message histories, adversaries can inject privileged instructions that bypass developer controls and increase attack success by up to 5×. While prior research assumed secure implementations; I showed that application-level flaws undermine those assumptions, leading to an underestimation of risk. This study exemplifies my measurement-first approach: examining examining live deployments exposes failure modes invisible to controlled evaluations. Following responsible disclosure, two major plugins issued security patches affecting thousands of websites.

Building Adaptable AI Systems

Most prior research prepares AI systems for foreseeable threats through methods like adversarial training. However, my work shows that these methods introduce hidden trade-offs, and deployment reveals a long tail of unpredictable failures. I translate these lessons into a new paradigm: designing *adaptable* AI systems that improve through deployment feedback. This vision has been the primary focus of my postdoc.

Addressing policy discrepancy in cyber defenses [14]. AI-based cyber defenses are typically trained on historical samples labeled as malicious or benign to recognize future threats. However, discussions with practitioners revealed a critical gap: labels are often context-dependent rather than inherent properties. For instance, security teams may mark legitimate alerts as false positives for business or operational reasons, and prior evidence suggests that most "false positives" arise this way [29]. My research formalizes this gap as *policy discrepancy* and develops practical methods to bridge it. We train models that can quickly adapt to an organization's current policy using a small set of labeled samples from deployment. This yields defenses that avoid predicting inherently ambiguous conditions in advance and instead adapt to them. To expand this direction, I led a successful Amazon Research Award proposal that now supports an ongoing collaboration.

LLM agents for security testing [15, 16]. LLMs show promise for automating security analysis but frequently hallucinate, producing incomplete or incorrect findings. During my postdoc, I worked with junior PhD students to develop LLM agents that systematically test PDF readers [15] and network protocols [16] for vulnerabilities. These systems discovered multiple zero-day vulnerabilities that prompted vendor patches. The key idea is to turn LLMs from static predictors into *adaptive testers*: they iteratively refine hypotheses, verify outcomes, retrieve supporting evidence, and learn from failed attempts through feedback loops.

Future Research Directions

AI is transforming computer security, offering immense promise while introducing new risks. My research 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 that develops *adaptive and antifragile* AI: systems that not only withstand real-world uncertainty but grow stronger from it. I will pursue this vision through two complementary directions.

Designing AI Systems for End-to-End Security

AI rarely operates in isolation. In practice, systems like malware detectors interact with rule-based filters, heuristic checks, and human analysts. Yet most research evaluates AI components in isolation, overlooking how they integrate into the broader defense. I aim to close this gap through a holistic approach that embeds AI into end-to-end security systems while addressing data limitations plaguing research.

Optimizing end-to-end systems. Real deployments must balance competing objectives, including accuracy, latency, scalability, transparency, and cost. I will extend the *learning with privileged information* paradigm [30] to train models using features available only during training (e.g., detailed program traces) while remaining efficient at deployment. Beyond individual models, I will co-train AI and non-AI modules to decide when to defer to one another, how to calibrate confidence, and how to optimize the trade-offs between speed and accuracy across the pipeline. This reframes robustness as an emergent property of the *entire defense system*, not of any single model.

Bridging the data gap. Industry holds realistic data but limited research capacity, while academia has researchers but not the data. My sandbox-to-endpoint study [2] exposed this divide: real-world data revealed failures invisible in academic benchmarks but could not be publicly shared. To reconcile this, I will develop *synthetic dataset generators* that reproduce key statistical and operational properties of proprietary data while protecting confidentiality. This approach will enable safe data sharing, reproducible research, and a collaborative ecosystem where academia and industry can jointly advance AI security.

Building Antifragile AI Agents Under Ambiguity

As AI agents take on open-ended tasks, they increasingly operate under underspecified or conflicting human goals. Robustness alone is not enough; future systems must be *antifragile*, improving through exposure to ambiguity rather than merely tolerating it. My work on *policy discrepancy* laid the foundation for this vision by showing that AI-based cyber defenses can adapt to new contexts that were never fully defined during training. The next step is to generalize this idea: enabling AI to learn which uncertainties matter in practice and refine its own behaviors through deployment feedback.

AI-assisted development as a testbed. Antifragility is becoming essential as software development shifts from code to natural language specifications. Security risks are shifting from classical vulnerabilities like memory-safety errors to logic flaws caused by misaligned assumptions between human intent and AI-generated implementations. Consider an e-commerce coupon system: are coupons single-use or reusable? If the requirement is unspecified, the AI must infer intent from training data, potentially introducing vulnerabilities. To optimize user experience, most AI systems favor silent inference over clarification requests, inadvertently amplifying such risks. Like policy shift, these challenges highlight a new frontier where system failures stem from *semantic ambiguity* rather than implementation error.

A unified framework for operating under uncertainty. I plan to develop a framework for antifragile AI that (1) measures how AI systems make assumptions under underspecification, (2) surfaces critical assumptions for human review, and (3) verifies alignment before deployment. Through iterative clarification, AI systems will learn which ambiguities truly matter, turning uncertainty into a source of improvement. Applications span code generation, autonomous decision-making, and cyber defense; domains where AI must act on incomplete human intent. Ultimately, I seek to understand AI not only as a target of vulnerabilities but as a source of new failure modes that demand a new generation of secure, adaptive, and self-correcting 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: *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 (NeurIPS)*. 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 (AISec)*. 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] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. “Fairness Without Demographics in Repeated Loss Minimization”. In: *35th International Conference on Machine Learning (ICML)*. 2018.

[28] 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).

[29] Bushra A Alahmadi, Louise Axon, and Ivan Martinovic. “99% false positives: A qualitative study of SOC analysts’ perspectives on security alarms”. In: *31st USENIX Security Symposium*. 2022.

[30] 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.