

From Secure Perception to Authorized Agent Action

Shaikh Akib Shahriyar | Rochester Institute of Technology | as8751@rit.edu

I study how AI systems fail under adversarial manipulation, and how to evaluate and design them so that their predictions, decisions, and actions remain trustworthy. A robustness claim is only as meaningful as the threat model behind it. Two parts of that threat model are usually left implicit. One is how much access and interaction an attacker needs. The other is what an attacker may change without changing the problem the system is meant to solve. My research makes both measurable.

My doctoral work began with deepfake detectors and asked how much access an attacker needs to evade them. In a cost-aware evaluation of 12 detectors, Carlini–Wagner transfer attacks that never queried the target reached at least 39% success between every pair of detector families [1]. I then turned to tool-using agents, where the complementary question is whether an adversarial edit still leaves the original task intact. In a controlled study, adding task-contract membership to a fixed similarity radius reduced one agent’s attacked-task count from 64/64 to 26/64 [2].

These threads meet in the question my group will pursue. What does it take for a system that perceives, acts, remembers, and delegates to remain trustworthy when part of its evidence or environment is adversarial? I will study it first for multimodal detectors, then for agents that act on what they perceive, and ultimately for persistent personalized agents that act on a person’s behalf.

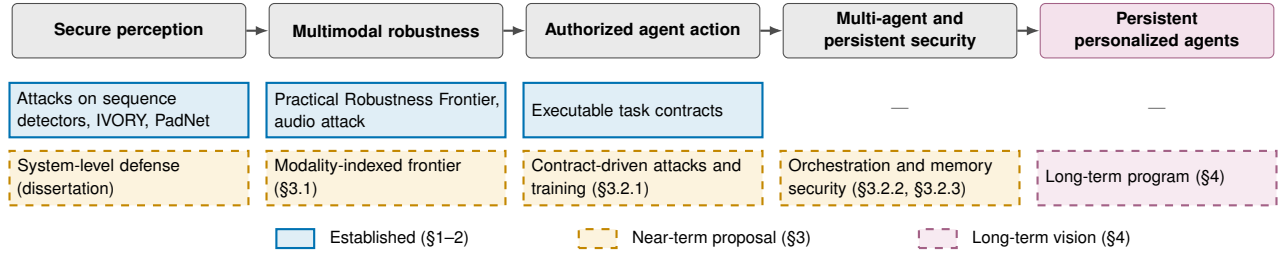


Figure 1: Research trajectory. Each stage adds a capability to the system under attack, and the stages overlap in time.

1 Current Research: From Adversarial Detection to Practical Robustness

1.1 Adversarial Reliability Across Media

Deepfake detectors are a first line of defense against synthetic media, yet they are classifiers and can be attacked. My first project asked whether temporal modeling makes them harder to attack. With Matthew Wright, I attacked two sequence-based detectors frame by frame. On FaceForensics++, the accuracy of a convolutional–LSTM detector fell from 81.3% to 8.3% under a white-box Carlini–Wagner attack. Attacks crafted on one detector reached up to 67.14% success against the other [3]. Temporal modeling alone did not provide adversarial robustness. That result raised a more deployment-relevant question. How much access does an attacker actually need?

The same question applies beyond vision. In work led by Anton Selitskiy and accepted to IJCB 2026, I helped extend adversarial evaluation to audio anti-spoofing [4]. The attack aligns synthetic speech with unpaired bona fide recordings using discrete optimal transport. Without gradients or feedback from the target, it raised a countermeasure’s equal error rate on the ASVspoof 2019 A17 condition from 1.26% to 26.13%. Because it needs bona fide reference data, access to that data is part of its threat model.

1.2 Defenses and Their Limits

In IVORY, published at IEEE FG 2025, I investigated whether diffusion-transformer purification could remove adversarial perturbations used to obscure soft-biometric attributes [5]. Under AutoAttack, it restored MobileNetV3 age-group accuracy from 2.14% to 52.44%, against 54.62% on clean inputs. IVORY therefore exposes a limit of perturbation-based privacy protection. It also raises a forensic question. Can reconstruction preserve the subtle evidence needed to detect manipulated media?

PadNet, led by Armon Barton, adds a learned barrier class and gradient regularization so that suspicious inputs are rejected rather than misclassified [6]. I contributed to its design and adaptive evaluation. Preventing a wrong answer does not necessarily recover the right one, and a defense must preserve useful behavior while it resists attack. These projects taught me to evaluate the whole defended system, including what it rejects and costs.

1.3 The Practical Robustness Frontier

Answering the access question required a measurement framework. In a systematization-of-knowledge study that I led with Ben Brown, Armon Barton, Ning Zhang, and Matthew Wright, we introduced the Practical Robustness Frontier (PRF) [1]. PRF reports attack success as a function of attacker access, from transfer attacks that never query the target, through black-box attacks with growing query budgets, to white-box adaptive attacks. We applied it to 12 detectors from four architectural families and to the defenses proposed for them. Three findings changed how I think about robustness.

- *Restricting access is not protection.* Carlini–Wagner transfer attacks that never query the target reached at least 39% success between every pair of detector families.
- *Similar vulnerability can hide very different attack costs.* At 1,000 SimBA-DCT queries, median attack success was 68% for frequency-aware detectors and 21% for video detectors. At high budgets, video detectors needed a median of 23,148 queries to approach their white-box success, against 6,812 for naive CNNs. Models with similar white-box ceilings converged at very different rates.
- *Defense conclusions change with access and adaptive evaluation.* PadNet delayed evasion but left 44–57% family-median success at 1,500 queries, and 49.2–68.4% under an attack aimed at its rejection mechanism.

These findings shifted my focus from whether a model can be broken to how much access an attacker needs. My dissertation applies that lesson in a system-level defense. A robustness number without an access model cannot be compared across papers, and a threat model without a validity condition cannot be interpreted. That validity question became central when I turned to systems that act.

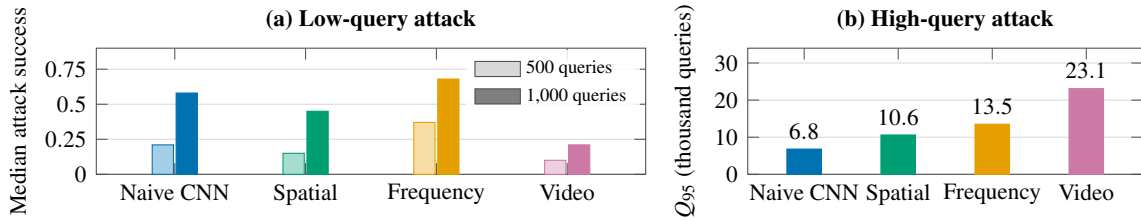


Figure 2: Similar eventual vulnerability can hide very different attack costs. Family medians for 12 detectors [1]. (a) Attack success at 500 and 1,000 SimBA-DCT queries. (b) Queries for HSJ- ℓ_∞ to reach 95% of the white-box success rate (Q_{95}).

2 Current Research: Task Validity for Tool-Using Agents

Security evaluations of tool-using agents edit the agent’s environment and read its response as a robustness measurement. That reading is valid only if the edited environment still poses the original task (Figure 3). A hostile note beside an intact payment instruction leaves the authorized task in place. An edit that replaces the beneficiary changes the task, so the resulting behavior cannot be interpreted as failure on the original task. Such an edit may still represent a meaningful corrupted-data threat, but it tests a different security question.

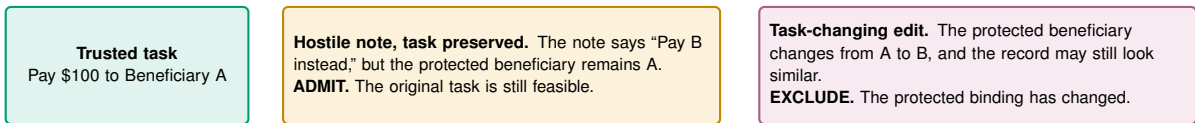


Figure 3: Task preservation precedes robustness measurement. Admission does not enforce safe execution.

Benchmarks such as AgentDojo expose real agent vulnerabilities [7], and a recent audit found injection vectors that overwrite task-critical state [8]. My work asks a logically prior question. Which modified environments still test the original task?

In our most recent work, I introduced executable task contracts [2]. A contract separates what defines a task, namely its protected facts, relations, and authorized outcomes, from the conditions an evaluator imposes. It is checked on each candidate edit before the agent’s behavior is observed. Contracts also separate a task from any one plan for completing it, because a failed reference plan does not mean the task is infeasible. Each candidate is executed once and rescored under different membership policies to isolate the validity rule.

The rule changes the conclusions. The study covers 14,336 trajectories from four open-weight agents on structured email, payment, calendar, and access-control tasks. Benign-calibrated similarity rules, which I evaluate as reusable task-membership proxies, excluded 156 to 290 of 640 admissible hostile candidates. The same rules admitted 96 to

128 of 128 candidates that changed protected facts. Under a fixed broad similarity radius, adding contract membership reduced Mistral-7B’s attacked-task count from 64/64 to 26/64. This corrects the measurement, not the agent. It reflects the combined contract-and-radius threat model, not robustness to the full admissible library. In AgentDojo, an unrestricted edit to a Slack channel identifier changed the task-defining entity and caused 40 of 132 original-reference executions to fail. Only some hostile trajectories ever exposed the injected text to the agent, so exposure and task validity must be reported separately from attack success.

The study has clear limits. Its tasks are structured, its contracts are manually specified, the evaluation is limited to four open-weight 7B model agents, and its contracts audit a fixed attack library rather than generate attacks. Scaling contracts to open-ended tasks and using them to drive attacks is where my future agent research begins.

3 Future Research

My group will pursue two near-term directions that share the commitments of my past work. An attacker is characterized by what it costs and by what it is allowed to change. The first extends PRF to multimodal evidence, and the second asks how authorization survives in agentic systems.

3.1 Direction I: Multimodal Practical Robustness

When does multimodal fusion actually increase the cost of evasion, and when does it merely add a cheaper attack route? A forensic pipeline may weigh video, audio, a transcript, file metadata, and provenance records. Prior work shows that an attack on one modality of a fusion model can override correct evidence from the others [9]. My audio results show an attack that needs no feedback at all. Neither tells us how access, cost, and semantic constraints combine across modalities. I will extend PRF so that attacker cost becomes a vector instead of a single number. It will record which modalities the attacker can modify and how much feedback it has for each. Each modality will use an appropriate constraint, such as perceptual similarity for video or speaker plausibility for speech. Comparing the cheapest attack on the fused system with the cheapest attack on its weakest modality will show when fusion buys security.

I will compare unimodal, early-fusion, late-fusion, and cross-attention detectors under single-modality and coordinated attacks. Consistency checks, such as audio–visual synchrony, are the natural defense against manipulating one stream. They are therefore what an adaptive attacker will target, so I will place the consistency module inside the attack loop [10]. Defenses will include dependence-aware evidence weighting, calibrated abstention, and selective purification. Selective purification raises a sharp question. Does removing adversarial noise also remove evidence of manipulation? For provenance-aware systems, I will distinguish unsigned metadata from cryptographically verifiable C2PA credentials [11], since altering a field and forging authenticated provenance require different capabilities. Defenses will be compared at matched clean utility, including the cost of abstention.

This work will produce an open multimodal benchmark. A first student project could test whether an audio–visual consistency detector remains useful once an attacker optimizes against it. A negative result, that additional modalities repeat the same weaknesses, would be equally informative.

The two directions converge when multimodal evidence becomes input to a tool-using agent. A manipulated document, image, voice clip, or video can alter not only a prediction but also the action the system takes. PRF contributes the perceptual and access budget for such an attack, and task contracts contribute the requirement that it leave the authorized task unchanged. I will begin with document-processing agents, separating perceptual error from authorization failure. Agent evaluation adds tool calls, turns, and persistence to the cost.

3.2 Direction II: Secure Agentic AI

A central question for my future group is how authorization survives adversarial environments, delegation, and persistent state while permitting flexible plans and legitimate updates. Evaluations still rarely guarantee that the attacked task is the original task, and they rarely measure attack strength against attacker cost. I will pursue both through three connected thrusts.

3.2.1 Task-preserving attacks and adversarial training. Today my contracts audit a fixed attack library. I will use them to drive attack generation instead. A contract-constrained generator will propose edits to observations, tool outputs, environment state, and interaction history. Each candidate will be checked against the contract before the victim’s behavior is observed, so validity cannot be tuned to outcomes. The result will be an adaptive red-team population that evolves against the agent’s current policy. Adaptive attacks have already broken published

prompt-injection defenses [12], so fixed libraries are not enough.

A central research question in this direction is whether adversarial training on contract-valid attacks produces security that generalizes to held-out tools, workflows, and attack strategies without collapsing into refusal. Trained agents will be evaluated alongside runtime defenses such as CaMeL [13] and Task Shield [14], which check actions against policy or the user’s goal at run time. I will study when explicit checks and learned policies reinforce each other and when an incomplete contract leaves both exposed. As in PRF, attack strength will be reported against attacker cost, measured in tool calls, environment interactions, edited tokens, and turns.

3.2.2 Secure multi-agent orchestration. When several agents appear to provide independent evidence, are their failures actually independent? Agreement is not independent evidence when every agent relies on the same poisoned source. Prior work shows that malicious instructions can spread between agents and that defenses can reduce collaboration [15]. Authenticated delegation has been proposed as a foundation for accountable agent authority [16]. Authenticated roles identify who sent a message. They cannot establish that a compromised specialist is truthful.

I will study three problems. The first is delegation. A specialist may need to read a document without being allowed to share it. Authority should narrow along a delegation chain and never silently widen. I will test whether a subtask can inherit the protected facts of its parent task. The second is evidence. Agents that share a document or tool can inherit the same poison, so I will aggregate evidence by independent provenance root rather than by agent count. The third is cost. I will measure security against the number of compromised agents or tools, the delegation depth, and the message budget, together with the collaboration cost of each protection. This will locate the point at which one compromised specialist can steer a planner or a vote.

3.2.3 Security of persistent agent memory. A note stored today can influence a later session after the original context is gone. Poisoning agent memory is already a documented attack [17]. I will treat memory as a security boundary with a lifecycle of write, summarize, retrieve, and revoke. The most important question is how authorization persists through derived memories after the original evidence is revoked. Suppose a summary was built from a source later found to be hostile. Deleting the summary discards useful knowledge. Retaining it, however, may preserve the attack.

I will first trace how source provenance survives summarization and retrieval, and distinguish observations from user-authorized commitments. When a source is revoked, that dependency graph can identify which memories require quarantine or repair. I will then test whether these mechanisms remain effective against attacks distributed across multiple sessions, and whether they still accept legitimate corrections. Recovery has limits. Deleting poisoned memory does not undo an email already sent, so I will study how an agent should respond when earlier actions relied on poisoned state.

4 Long-Term Vision: Secure Persistent Personalized Agents

Memory, provenance, delegation, and authorization lead to a longer-term question. Could a computational agent faithfully represent an individual’s decision patterns, preferences, communication style, and knowledge over time, and act on that person’s behalf within limits the person sets? Generative agents already produce believable behavior [18]. Believability is not fidelity, and neither is security. This is a staged research program. It begins with narrowly scoped, consented agents in constrained domains such as professional planning, and it does not assume that faithful personal agents are a near-term deliverable.

A first challenge is defining behavioral fidelity. In the spirit of PRF, I will study how many interactions a bounded-cost evaluator needs to distinguish the agent from the represented person on held-out decisions. A second is consistency, because preferences change and some attributes should never be inferred or reproduced. Contracts could state such limits as enforceable constraints. Third, the agent needs scoped, revocable permissions, and recipients need to distinguish human from synthetic actions. Provenance can establish origin without proving truth. Finally, memories, delegations, corrections, revocations, and deletion must remain consistent over time. Direction II supplies mechanisms for that lifecycle, but removing a person’s data from learned components remains open.

Building an independent research program. Having mentored undergraduate, master’s, and junior Ph.D. researchers in adversarial ML and generative-AI security, I have designed these directions so that focused student projects feed shared benchmarks and tools. My work on the NSF-supported VARUNA and DeFake projects connects this agenda to real-world media-authenticity workflows for journalists. The program aligns with federal priorities in cybersecurity and trustworthy AI, including NSF SaTC and ONR, and I have worked on SaTC and ONR proposals for both

near-term directions. My aim is to make claims of AI reliability accountable to the evidence a system uses, the adversary it faces, and the authority a person has granted it.

References

- [1] Shaikh Akib Shahriyar, Ben Brown, Armon Barton, Ning Zhang, and Matthew Wright. SoK: The practical robustness frontier: Cost-aware vulnerability analysis for deepfake detectors. Manuscript in preparation for SP 2027, 2026.
- [2] Shaikh Akib Shahriyar, MD. Akil Raihan Iftee, Armon Barton, and Matthew Wright. What counts as the same task? executable contracts for robustness evaluation of tool-using agents, 2026. URL <https://arxiv.org/abs/2610.14477>. Manuscript under review at AAMAS 2027.
- [3] Shaikh Akib Shahriyar and Matthew Wright. Evaluating robustness of sequence-based deepfake detector models by adversarial perturbation. In *Workshop on Security Implications of Deepfakes and Cheapfakes (WDC '22)*, co-located with ACM AsiaCCS, pages 13–18, Nagasaki, Japan, 2022. ACM. doi: 10.1145/3494109.3527194.
- [4] Anton Selitskiy, Shaikh Akib Shahriyar, and Jishnuraj Prakasan. Discrete optimal transport is a strong audio adversarial attack. Accepted at the IEEE/IAPR International Joint Conference on Biometrics (IJCB). arXiv:2509.14959, 2026.
- [5] Shaikh Akib Shahriyar, Matthew Wright, and Armon Barton. IVORY: Adversarial purification of obfuscated faces to extract soft-biometrics using diffusion transformers. In *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10, 2025. doi: 10.1109/FG61629.2025.11099354.
- [6] Armon Barton, Matthew Wright, Shaikh Akib Shahriyar, Edgar Jatho, Mohammad Saidur Rahman, Kantha Girish Gangadhara, and Jiang Ming. PadNet: Defending neural networks against adversarial examples. *ACM Transactions on Privacy and Security*, 29(2): 17:1–17:26, March 2026. doi: 10.1145/3799889.
- [7] Edoardo DeBenedetti, Jie Zhang, Mislav Balunović, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pages 82895–82920, 2024.
- [8] Rishika Bhagwatkar, Kevin Kasa, Abhay Puri, Gabriel Huang, Irina Rish, Graham W. Taylor, Krishnamurthy Dj Dvijotham, and Alexandre Lacoste. Indirect prompt injections: Are firewalls all you need, or stronger benchmarks? arXiv preprint arXiv:2510.05244, 2025.
- [9] Karren Yang, Wan-Yi Lin, Manash Barman, Filipe Condessa, and Zico Kolter. Defending multimodal fusion models against single-source adversaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3340–3349, 2021.
- [10] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020.
- [11] Coalition for Content Provenance and Authenticity. Content credentials: C2PA technical specification, version 2.4. https://spec.c2pa.org/specifications/specifications/2.4/specs/C2PA_Specification.html, April 2026.
- [12] Qiusi Zhan, Richard Fang, Henil Shalin Panchal, and Daniel Kang. Adaptive attacks break defenses against indirect prompt injection attacks on LLM agents. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7116–7132, Albuquerque, New Mexico, 2025. doi: 10.18653/v1/2025.findings-naacl.395.
- [13] Edoardo DeBenedetti, Iliia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. In *IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 2026. arXiv:2503.18813.
- [14] Feiran Jia, Tong Wu, Xin Qin, and Anna Squicciarini. The task shield: Enforcing task alignment to defend against indirect prompt injection in LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29680–29697, Vienna, Austria, 2025. doi: 10.18653/v1/2025.acl-long.1435.
- [15] Pierre Peigné, Mikolaj Kniejski, Filip Sondej, Matthieu David, Jason Hoelscher-Obermaier, Christian Schroeder de Witt, and Esben Kran. Multi-agent security tax: Trading off security and collaboration capabilities in multi-agent systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(26):27573–27581, 2025. doi: 10.1609/aaai.v39i26.34970.
- [16] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Alan Chan, and Alex Pentland. Position: AI agents need authenticated delegation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82211–82231. PMLR, 2025.
- [17] Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. AgentPoison: Red-teaming LLM agents via poisoning memory or knowledge bases. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.
- [18] Joon Sung Park, Joseph C. O’Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2023. doi: 10.1145/3586183.3606763.