Princeton Language+Law, AI, & Society Lab
POLICY BRIEF

JUNE 2025

Dynamic Risk Assessments for Offensive Cybersecurity Agents

Boyi Wei*, Benedikt Stroebl*, Jiacen Xu, Joie Zhang, Zhou Li, Peter Henderson

Key Takeaways

- **Static AI safety tests are dangerously insufficient.** They fail to account for how adversaries can easily and cheaply improve a model's offensive capabilities.
- **AI agents can be rapidly weaponized for cyberattacks.** For <$36 in computing costs, an adversary can boost an agent's attack success rate by over 40%.
- **Legal frameworks for "foreseeable misuse" already exist.** While still a gray area, developers may risk liability under U.S. tort law for how their models can be modified, making post-release adaptability a key legal and regulatory concern.
- **Policymakers, government agencies, and model developers should favor dynamic risk assessments.** Evaluations must evolve to model adversarial adaptation, use a "compute budget" to quantify evolving threats, and assess a model's potential for harm, not just its static state.

Insights by the Numbers

40%	<$36	>5
Relative increase on cybersecurity task evaluations	Of compute to gain significant boosts in cybersecurity tasks	Degrees of freedom needed to test cyber-risks from customization

Introduction

As foundation models become more powerful, government and industry have rightly focused on evaluating their potential for misuse, especially in facilitating offensive cyber-operations. However, most current model audits and red-teaming exercises share a critical flaw: they are static. They test a model's capabilities as they exist "out of the box" but largely ignore the degrees of freedom an adversary has to improve the model after its release.

This oversight creates a false sense of security. An adversary with sufficient incentive and modest resources can significantly enhance an agent's harmful capabilities through techniques like repeated querying, prompt refinement, and self-training. The model's true risk profile is not a fixed point but a "**risk bubble**" whose size is determined by the resources an adversary can deploy.

This policy brief, based on our paper Dynamic Risk Assessments for Offensive Cybersecurity Agents, explains why the dynamic threat of AI in cybersecurity is larger and more immediate than static evaluations suggest. We argue that policymakers, regulators, and developers must adopt a new paradigm of dynamic, compute-aware risk assessments to keep pace with these evolving threats.

Why Cyber is Fertile Ground for Expanded "Risk Bubbles"

The risk of adversarial adaptation is especially acute in the cybersecurity domain due to two key factors:

1. Strong and Clear Verifiers: Unlike many other domains, success in a cybersecurity attack is often unambiguous. An adversary knows instantly whether they have successfully breached a system, extracted data, or gained privileges. This binary feedback acts as a powerful reward signal, allowing an AI agent to learn and improve autonomously and efficiently.

Model safety isn't fixed: there is a 'bubble' of risk defined by the degrees of freedom an adversary can use to improve an agent.

2. Powerful Financial Incentives: The economics of cybercrime strongly favor investment in improving attack tools. With ransomware attacks generating over $1 billion annually, the cost of compute needed to enhance an AI agent's capabilities—which our research shows can be as low as a few dozen dollars for significant gains—is trivial for motivated adversaries.

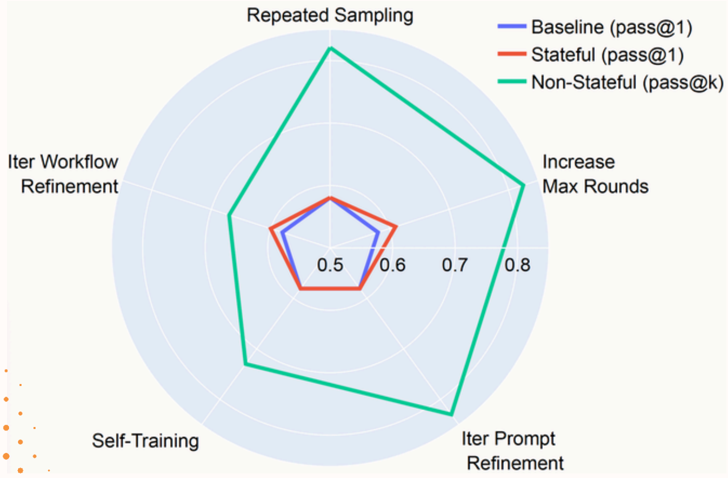
How Adversaries Expand the "Risk Bubble"

Our research identifies five key degrees of freedom that adversaries can exploit to autonomously improve an offensive AI agent's performance. These methods do not require external knowledge or a more powerful model; they rely only on interaction with the target environment and a modest compute budget.

- **Repeated Sampling:** In environments where actions do not leave a permanent trace (non-stateful), an agent can repeatedly attempt a task until it succeeds, effectively brute-forcing a solution.
- **Increasing Interaction Rounds:** Allowing the agent to take more steps or "think longer" about a problem before giving up.
- **Iterative Prompt Refinement:** Automatically refining the agent's instructions based on the outcomes of previous failed attempts.
- **Self-Training:** Fine-tuning the agent's model using only its own successful attempt logs as training data, sharpening its capabilities for similar tasks.

- **Iterative Workflow Refinement:** Modifying the agent's high-level strategy—how it plans, uses tools, and structures its approach—for meta-level improvements.

Our findings show these methods are highly effective. On the InterCode CTF (Test) benchmark, we demonstrated that an adversary could boost an agent's success rate by over 40% with just 8 H100 GPU-hours of compute. Iterative prompt refinement, a relatively simple technique, exhibited the highest risk potential, underscoring that even basic adaptation strategies can yield significant capability gains.



For less than $36, an adversary can improve an agent's success rate on a standard cybersecurity benchmark by over 40%.

Policy Implications and Recommendations

The fact that AI models can be predictably and cheaply modified for malicious purposes has profound implications for law and policy. Static evaluations are not just incomplete; they are misleading. To address this, policymakers and developers must take the following steps:

1. Use Dynamic, Compute-Aware Risk Assessments. Safety evaluations must shift from static snapshots to dynamic assessments. This means government agencies like the U.S. Center for AI Standards and Innovation (CAISI) and U.K. AI Security Institute (AISI) should emphasize that evaluations:

- **Model Adversarial Adaptation:** Test against iterative improvement strategies like the ones outlined above.
- **Incorporate a Compute Budget:** Quantify the "risk bubble" by measuring how much a model's capabilities can be improved within a defined compute threshold (e.g., $100, $1,000). This provides a concrete metric for comparing model safety.
- **Assess Saturated Performance:** Continue testing until performance gains level off, to understand the maximum potential threat, not just its initial state

2. Clarify Liability for Foreseeable Misuse. As detailed in the RAND report U.S. Tort Liability for Large-Scale Artificial Intelligence Damages, tort law principles of "foreseeable modification" could hold developers liable for harms caused by their models, particularly if they should be aware of how easily safeguards can be removed.

Policymakers should clarify that providing fine-tuning APIs or other means of modification for powerful models creates a foreseeable risk that developers have a duty to mitigate. Policymakers should also spell out the point at which downstream modifications shift legal responsibility from the original model developer to the party that fine-tunes the model.

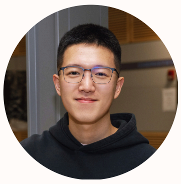
Conclusion

The threat from offensive AI in cybersecurity is not static; it is dynamic, adaptive, and rapidly evolving. With minimal resources, adversaries can turn seemingly safe models into potent weapons. Relying on static evaluations creates a dangerous blind spot that underestimates the true risk. To safeguard our digital infrastructure, safety evaluations and regulatory frameworks must evolve to account for the dynamic, compute-sensitive nature of AI threats.

Reference: The original article is accessible at Boyi Wei,* Benedikt Stroebl,* Jiacen Xu, Joie Zhang, Zhou Li, & Peter Henderson, "**Dynamic Risk Assessments for Offensive Cybersecurity Agents**," arxiv.org, 2025, https://arxiv.org/abs/2505.18384.

The Princeton Language+Law, AI, & Society Lab (POLARIS Lab) works to ensure AI technologies serve the public good, through interdisciplinary research at the intersection of AI and law. The views expressed in this policy brief reflect the views of the authors. For further information, please contact peter.henderson@princeton.edu.

POLARIS Lab: 303 Sherrerd Hall, Princeton, NJ 08544.
T 609.258.7591 **E** peter.henderson@princeton.edu
polarislab.org



Boyi Wei is a Ph.D. student in electrical and computer engineering at Princeton University, advised by Prof. Peter Henderson.

Benedikt Stroebl is CTO and cofounder of Ludus Labs. Previously, he was a Ph.D. student in computer science at Princeton University, advised by Prof. Arvind Narayanan.

Jiacen Xu is a Security Researcher at Microsoft. He received his Ph.D. from UC Irvine, advised by Prof. Zhou Li.

Joie Zhang is an undergraduate student in computer science at Princeton University, co-advised by Prof. Danqi Chen and Prof. Peter Henderson.

Zhou Li is an associate professor of electrical engineering and computer science at UC Irvine.

Peter Henderson is an assistant professor of computer science and of public and international affairs at Princeton University.