

RestriktAI: Enhancing Safety and Control for Autonomous AI Agents

João Lucas Duim Juan Belieni
impactRIO *impactRIO*

This proposal addresses a critical gap in AI safety by mitigating the risks posed by autonomous AI agents. These systems often require access to sensitive resources that expose them to vulnerabilities, misuse, or exploitation. Current AI solutions lack robust mechanisms for enforcing granular access controls or evaluating the safety of AI-generated scripts. We propose a comprehensive solution that confines scripts to predefined, sandboxed environments with strict operational boundaries, ensuring controlled and secure interactions with system resources. An integrated auditor LLM also evaluates scripts for potential vulnerabilities or malicious intent before execution, adding a critical layer of safety. Our solution utilizes a scalable, cloud-based infrastructure that adapts to diverse enterprise use cases.

1 Problem overview

As autonomous AI systems become increasingly integrated into various workflows, there is a rising demand for these systems to function as autonomous agents. This development presents several risks, including potential misuse by malicious actors and vulnerabilities that may arise from untrusted or misaligned models.

Many tasks involve access to sensitive resources, such as file systems, databases, and the internet. However, current AI systems frequently lack the capability to enforce granular access controls, which could result in exploitation of resources, unauthorized actions, or unintended consequences.

Malicious actors could potentially exploit AI systems to generate or execute scripts intended to compromise systems or extract data. This concern is particularly pronounced when AI agents are granted unrestricted scripting capabilities without appropriate oversight. AI-generated scripts may inadvertently introduce vulnerabilities. In the absence of mechanisms to assess and mitigate these vulnerabilities, the risk of exploitation significantly increases.

2 Solution

Technical foundation

- **Granular permission system:** AI agents are able to generate scripts that operate only within strictly controlled environments. Scripts will have access only to predefined resources using techniques akin to sandboxing. For example, a script generated by the AI agent could be configured to read data only from a specific directory, without access to any other system resources or features.
- **Vulnerability assessment via LLMs:** A auditor LLM is incorporated to evaluate and flag potentially malicious scripts before execution. This technique is proven to vastly increase the safety of AI systems without compromising the usefulness of the model [1].

- **Simplified script language:** The scripts are written in a lightweight, structured scripting language optimized for easy generation and parsing by AI systems. This language ensures precise control and predictability over script behavior, while also enabling structured generation.

How this solution addresses the problem?

- By isolating scripts with granular permissions, the system prevents unaligned or malicious actions that could exploit broader system resources.
- The vulnerability assessment layer provides an additional security checkpoint, minimizing risks of deceptive or harmful operations in the environment.

Competitive advantages

- The combination of permission-controlled scripting and LLM-based vulnerability assessments provides a multi-layered approach to AI agents safety, addressing both prevention and detection of risks. To the best of our knowledge, existing companies in the market are not addressing this gap.
- Developing a proprietary scripting language tailored for AI systems could be patented, offering a unique value proposition for enterprises seeking secure AI-driven automation.
- The system proactively addresses emerging regulatory requirements around AI safety and accountability by ensuring strict operational boundaries for AI-generated scripts, offering businesses a compliant and future-proof approach to adopting AI systems.
- The system's fine-grained control over AI actions makes it uniquely suited for high-risk tasks that require strict oversight, such as financial market analysis, critical infrastructure automation, or security-related operations. By enforcing granular permissions and employing robust vulnerability assessment, it ensures that AI agents can only interact with approved resources, preventing unintended or malicious actions.

Innovation and scalability

- The solution integrates AI safety principles into practical AI deployment, innovating by combining structural (permissions) and behavioral (vulnerability detection) safeguards.
- The cloud-based infrastructure ensures scalability, enabling enterprises to adopt this solution seamlessly for diverse tasks. The lightweight scripting language also ensures high performance and adaptability across use cases.

3 Demo

Objective

In this demo, we examine the scenario in which a malicious user prompts the agent to perform an SQL injection on a website. The details are as follows:

- A user submits a prompt designed to generate the malicious script, e.g., "Perform an SQL injection test on example.com."
- The agent will likely deny the generation for safety reasons but, in rare cases, may generate the script based on the prompt.

- If the script is generated, the auditor will flag the code as malicious and stop the execution of the script.

Technical details

- We load the QWEN/QWEN2.5-7B-INSTRUCT model [2] as an agent using the TRANSFORMERS library [3]. The agent specification is tailored to generate Python scripts, so we reimplemented the `python_evaluator` method for vulnerability assessment.
- The granular permission system is very compatible with the TRANSFORMERS library. By default, the agent has access to a limited set of built-in Python libraries, and it can be easily extended. For safety considerations, we did not supply the necessary libraries that would enable the agent to execute malicious actions.
- The code for this demo is available at <https://gist.github.com/juanbelieni/8921b1d7957db6bf15a0f0f87e481d40>.

Results

The agent successfully identifies and rejects the prompt crafted to produce a SQL injection script most of the time. In the instances when we mislead the model into believing the website belonged to one of us, the modified `python_evaluator` recognized the output as malicious and prevented execution, thus demonstrating the effectiveness of the vulnerability assessment framework. This underscores the efficacy of implementing safety-oriented model configurations and runtime auditing to mitigate risks associated with malicious user inputs.

4 Process

Timeframe	What will you do?
Next 3 months	• Hire people specialized in AI safety, LLMs, and secure system design to research and enhance AI agent capabilities and safety. • Develop a beta version of the system, including the script generation framework, permission sandboxing, and vulnerability assessment module. • Build a dedicated platform for AI safety researchers and selected individuals to test the system and perform red-teaming.
2025	• Launch the platform publicly and scale its availability for companies seeking secure AI-driven solutions. • Invest in marketing campaigns to raise awareness about the platform's security features and regulatory compliance benefits. • Develop and execute a strategic lobbying initiative to advocate for regulations promoting secure automation practices, creating a framework where AI labs will benefit from integrating our product to meet compliance requirements.

2026	• Launch a developer ecosystem, including an open API and SDKs, to encourage third-party developers to build applications and services on the platform. • Expand into other domains, such as data analysis for financial market companies, leveraging the platform’s security and automation capabilities. • Establish partnerships with major AI labs to enhance research collaboration and integrate cutting-edge technologies.
2027	• Host an annual global summit focused on AI safety, innovation, and ethical deployment of AI agents, bringing together stakeholders from industry, government, and academia to solidify the platform’s leadership in the domain • Advocate for the creation and adoption of international standards for the deployment of secure and ethical AI agents, positioning the platform as the global leader.

5 Impact on AI safety & key risks

Our solution directly addresses the risks posed by malicious users of agentic AI models intending to jailbreak them and the risks posed by misaligned AI agents, especially when suffering distribution shifts. Moreover, we expect to indirectly contribute to the advancement of AI safety regulations in key countries, to the establishment of global safety standards of AI agents and to increase the public awareness regarding the risks of using agentic AIs without the appropriate limitations and oversight.

When scaling oversight using an auditor LLM, the risk of such auditor being deceived increases. To mitigate this, we intend to concentrate our team’s efforts by constantly red-teaming the auditor model and by performing cutting-edge research in collaboration with academia and other organizations. Another main risk appears when AI agents are powerful enough to find a way out of the box, breaking out the limited environment. It’s not clear if this can be avoided at all, but we would invest in research to develop an environment mathematically proved to be impossible to cheat. Note that, to our knowledge, no risks would counterfactually increase due to RestriktAI’s existence.

Currently, there are isolated efforts to make AI agents safer. Since this technology is yet in its infancy, the risks scenarios could change drastically in the near future, so the safety industry needs new frameworks that can escalate at least as fast. For that reason, we believe RestriktAI is innovatively well aligned with such industry and will push the boundaries of safety standards.

We believe in adaptive optimized balance between automated and human oversight to make monitoring and evaluation effective. The fast pace of agents development will require great flexibility to rapidly adapt our strategies. Ideally, the team would always spot minimal evidence of possible flaws in the platform and immediately solve it. In extreme cases, all our virtual environments could be plugged off to avoid any leakage.

Initially, a great metric would be the sum of “unsafe” or “access denied” flags raised by our system, because each of those requests could have caused some amount of harm if RestriktAI hadn’t been founded. However, cases when such flags are wrongly raised should be taken into account. In addition, each AI lab that becomes compliant to integrate our product into every AI agent they develop (due to regulations or deliberate collaboration) would count to our KPI, bonus points according to the lab’s relevance.

References

- [1] Ryan Greenblatt et al. *AI Control: Improving Safety Despite Intentional Subversion*. 2024. arXiv: 2312.06942 [cs.LG]. URL: <https://arxiv.org/abs/2312.06942>.
- [2] Qwen Team. *Qwen2.5: A Party of Foundation Models*. Sept. 2024. URL: <https://qwenlm.github.io/blog/qwen2.5/>.
- [3] Thomas Wolf et al. “Transformers: State-of-the-Art Natural Language Processing”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. URL: <https://aclanthology.org/2020.emnlp-demos.6/>.