Politecnico di Torino

COMPUTER AND CONTROL ENGINEERING

Istituto AI4I - Formal Verification of Alignment in Autonomous AI-Agents

Funded By	Istituto Italiano sull'Intelligenza Artificiale per l'Industria [P.iva/ CF:13130030011]
Supervisor	SISTO RICCARDO - riccardo.sisto@polito.it
Contact	NICOLA FRANCO (nicola.franco@ai4i.it)
Context of the research activity	Autonomous AI agents are increasingly deployed in web-based environments. However, they face challenges: maintaining alignment with human objectives while operating securely in adversarial environments that may attempt to manipulate their behavior or compromise data integrity. The research aims to develop formal verification methods to provide value alignment and security robustness guarantees. The activities will be carried out in a collaboration between the AI4I and Politecnico di Torino.
	The deployment of autonomous AI agents in web environments presents critical challenges that have not been adequately addressed by current verification approaches. While existing work focuses either on AI alignment or cybersecurity independently, the intersection of these concerns remains largely unexplored. Recent incidents have demonstrated that adversarial web content can manipulate agent behavior, leading to misalignment, data breaches, or unintended actions that violate user objectives. The main objective of the proposed research is to advance the state of the art in formal verification for autonomous AI agents by developing techniques that simultaneously guarantee value alignment and security robustness in adversarial web environments. This will be achieved by first establishing formal specifications for aligned agent behavior using temporal logic, then developing verification techniques that can provide both deterministic and statistical guarantees of alignment preservation under adversarial conditions. The research will produce AI agents with verified safety properties and demonstrate practical scalability for real-world applications. The research activity will be organized in three phases: Phase 1 (1st year): The candidate will conduct a comprehensive analysis of the threat landscape for web-based AI agents, identifying attack vectors that could lead to misalignment or security breaches. This includes studying prompt injection attacks, adversarial inputs through web content, data poisoning, and other manipulation techniques specific to web environments.

Objectives	The candidate will develop formal specifications using temporal logic (such as LTL or CTL) to capture alignment requirements and security properties for agents operating in e-commerce, web search, and automated browsing contexts. At this phase's end, preliminary results are expected to be published, including a taxonomy of adversarial threats to agent alignment, formal specification frameworks for aligned behavior, and initial case studies demonstrating vulnerability patterns. During the first year, the candidate will also acquire necessary background knowledge by attending courses in formal verification, AI safety, and adversarial machine learning, supplemented by personal study. Phase 2 (2nd year): The candidate will develop model checking techniques and verification algorithms specifically designed for autonomous AI agents. This phase focuses on creating methods that can verify temporal logic specifications against agent behavior models, accounting for both deterministic and probabilistic behaviors in adversarial settings. The candidate will design architectures with built-in verification-friendly properties, such as compositional structures or certified robustness guarantees, that facilitate formal analysis while maintaining practical performance. Special emphasis will be placed on statistical verification techniques that can provide probabilistic guarantees of alignment under uncertainty, addressing the challenge that complete formal verification may be computationally intractable for complex neural models. The candidate will implement prototype verification tools and evaluate them on realistic benchmarks from web browsing, e-commerce recommendation, and search agent domains. The results of this work will be submitted for publication, aiming at top-tier venues in formal methods, AI safety, or cybersecurity, with at least one conference/journal publication expected. Phase 3 (3rd year): Based on results from the previous phase, the candidate will address scalability challenges to enable practical deployment of the verification techniques. This includes developing efficient verification algorithms, creating automated specification synthesis methods, and establishing workflows for integrating verification into the agent development lifecycle. The candidate will conduct comprehensive experimental evaluations on real-world applications, demonstrating both the effectiveness of the verification approach in detecting misalignment vulnerabilities and its computational feasibility for practical systems. The work will produce open-source verification tools and benchmark datasets to facilitate adoption by practitioners and future research. The final phase will also complete dissemination activities, including publications, tool releases, and potentially technology transfer activities. The contributions produced by the proposed research can be published in conferences and journals belonging to the areas of Formal Methods (e.g., CAV, TACAS, FM, ACM Transactions on Software Engineering and Methodology), AI Safety and Machine Learning (e.g., NeurIPS, ICML, ICLR, AAAI, Journal of Artificial Intelligence Research), and Cybersecurity (e.g., IEEE S&P, ACM CCS, USENIX Security, NDSS, IEEE Transactions on Dependable and Secure Computing).
Skills and competencies	The candidate should have a strong background in at least two of the following areas: Formal methods/verification, artificial intelligence/machine learning, or cybersecurity.

competencies for the development of the activity	Familiarity with temporal logic, model checking, or neural network verification is welcome, as well as programming proficiency with AI frameworks (such as PyTorch or TensorFlow). The candidate can acquire the missing knowledge during the PhD Program, by exploiting courses and the expertise available in the research group.