COMPUTER AND CONTROL ENGINEERING

Istituto AI4I - Adversarial Robustness in Multi-Modal Foundation Models

Funded By	Istituto Italiano sull'Intelligenza Artificiale per l'Industria [P.iva/ CF:13130030011]
Supervisor	CAGLIERO LUCA - luca.cagliero@polito.it
Contact	GIORDANO DANILO - danilo.giordano@polito.it NICOLA FRANCO (nicola.franco@ai4i.it)
Context of the research activity	Multi-modal AI models that integrate vision, language, and audio processing are becoming increasingly prevalent in critical applications such as content moderation, customer service automation, and AI-assisted software development. However, these systems introduce novel attack surfaces arising from cross-modal interactions, where adversarial inputs in one modality can exploit semantic inconsistencies or vulnerabilities when processed jointly with other modalities. This research aims to systematically investigate these vulnerabilities and develop novel attack methods that expose weaknesses in cross-modal processing.
	Large-scale multi-modal AI models represent a significant advancement in AI, enabling richer understanding and generation of content across different sensory modalities. However, the security implications of cross-modal interactions remain poorly understood. Recent evidence suggests that adversaries can craft inputs where, for example, benign visual content paired with carefully manipulated audio or text can cause model misclassification or unsafe outputs. Unlike single-modality adversarial attacks, multi-modal attacks exploit the complex fusion mechanisms that integrate information across modalities, creating attack vectors that are difficult to detect and defend against using existing techniques. The main objective of the proposed research is to advance understanding of adversarial vulnerabilities in multi-modal AI systems by establishing theoretical foundations for attack surfaces and developing practical defense mechanisms that can be deployed in real-world applications. In this research work, the candidate will leverage expertise in adversarial machine learning, information theory, and secure system design to develop both theoretical insights and practical solutions for multi-modal AI security. The research activity will be organized in three phases: Phase 1 (1st year): The candidate will conduct a comprehensive study of the attack surface of multi-modal AI models, focusing on architectures that

Objectives	combine vision, language, and audio modalities. This phase involves analyzing state-of-the-art multi-modal fusion mechanisms (including early fusion, late fusion, and attention-based fusion) to identify potential vulnerability points where cross-modal interactions can be exploited. The candidate will develop a taxonomy of attack vectors specific to multi-modal systems, categorizing them by the exploited modality interactions, attack objectives (e.g., misclassification, content injection, behavior manipulation), and required adversary capabilities. The candidate will begin developing novel attack methods that exploit semantic inconsistencies between modalities. For example, attacks where visual and textual content appear individually benign but their combination triggers misclassification, or where subtle audio perturbations alter the interpretation of accompanying visual content. At this phase's end, preliminary results are expected to be published, including the attack taxonomy, initial attack methods, and theoretical characterizations of multi-modal vulnerability surfaces. During the first year, the candidate will also acquire necessary background through coursework in adversarial machine learning, information theory, and multi-modal AI architectures, supplemented by personal study. Phase 2 (2nd year): The candidate will create a dataset of synthetic injection attacks, systematically categorized by attack technique (e.g., cross-modal perturbation, semantic inconsistency exploitation, modality-specific backdoors), target vulnerability, and application context. This dataset will serve as a benchmark for evaluating defense mechanisms and will be made publicly available to support the research community. Experimental evaluation will demonstrate the effectiveness of the attack methods in exposing vulnerabilities. Applications in content analysis systems, customer service AI agents, and AI-assisted software development tools will serve as case studies. The results of this work will be submitted for publication, targeting top-tier venues in machine learning security and AI, with at least one publication expected. Phase 3 (3rd year): Building on previous results, the candidate will develop new attack techniques capable of automatically synthesizing attack samples. This library will implement parameterized attack templates that can generate diverse adversarial examples across different multi-modal architectures. The attack generation framework will incorporate the results from Phase 2 to prioritize attack strategies with higher success probabilities. The candidate will refine and validate the attack mechanisms through extensive testing on large-scale datasets. The final phase will produce comprehensive documentation, open-source implementations of attack tools, and best practices guidelines for developing robust multi-modal AI systems. Dissemination activities will include publications, software releases, and potentially workshops or tutorials to transfer knowledge to practitioners. The work may be conducted in collaboration with industry partners deploying multi-modal AI systems or relevant research initiatives focusing on AI safety and security, providing access to realistic deployment scenarios and validation opportunities. The contributions produced by the proposed research can be published in conferences and journals belonging to the areas of Machine Learning and AI (e.g., NeurIPS, ICML, ICLR, CVPR, ACL, AAAI, Journal of Machine Learning Research, IEEE Transactions on Pattern Analysis and Machine Intelligence), Cybersecurity (e.g., IEEE S&P, ACM CCS, USENIX Security, NDSS, ACM

	Transactions on Privacy and Security), and interdisciplinary venues focusing on AI safety (e.g., AIES, FAccT).

Skills and competencies for the development of the activity	To successfully develop the proposed activity, the candidate should have a strong background in machine learning and artificial intelligence, with particular emphasis on deep learning architectures. Familiarity with multi-modal AI models, computer vision, natural language processing, or audio processing is highly desirable. Strong programming skills and experience with deep learning frameworks (such as PyTorch or TensorFlow) are essential. The candidate can acquire specialized knowledge in adversarial robustness or specific multi-modal architectures as part of the PhD Program, by exploiting specialized courses and the research group's expertise.