Graduate School of Science and Technology Master's Thesis Abstract

Laboratory name (Supervisor)	Cyber Resilience (Youki Kadobayashi (Professor))		
Student ID	2311430	Submission date	2025 / 1 / 20
Name	VALERIE MEGAN		
Thesis title	Detecting Malicious AI-Generated Personas Through Hybrid Systems Using Rule-Based Logic and Machine Learning-Assisted Models		

Abstract

The rapid advancement of artificial intelligence (AI) has led to the emergence of AI-generated personas, digital entities designed to mimic human behavior with remarkable realism. While these personas offer innovative applications in marketing, entertainment, and customer service, they pose significant risks, including misuse in misinformation campaigns, identity theft, and social engineering attacks. This study addresses these challenges by proposing a hybrid detection framework integrating rule-based logic with machine learning-assisted models. A key contribution of this research is the development of a novel taxonomy of harmfulness, categorizing harmful personas based on intent, amplification, and consequences, which is complemented by the creation of a comprehensive dataset comprising benign and harmful personas for rigorous model testing. The proposed detection system adopts a holistic approach, evaluating personas as composite entities rather than solely focusing on their content. Experimental results demonstrate the effectiveness of the hybrid system in identifying both explicit and subtly harmful personas. These findings emphasize the necessity of holistic evaluation methods to address the evolving sophistication of personas. Furthermore, the research explores the ethical implications of personas and provides actionable recommendations for policymakers and industry stakeholders to mitigate associated risks. By advancing detection methodologies and emphasizing ethical considerations, this study contributes to the growing field of AI safety and lays the foundation for creating a safer and more trustworthy digital environment, while outlining pathways for future work, including enhancing detection algorithms, expanding datasets, and examining the societal impacts of personas through interdisciplinary research.