Center for Research on AI Security and Technology Evolution (CREATE)

OUTLINE

Center for Research on AI Security and Technology Evolution (CREATE) promotes research and development based on three pillars: Security for AI, AI for Security and Trustworthy AI.

CREATE aims to establish reliable and secure AI technologies that can respond to increasingly sophisticated cyberattacks and the rapid social adoption of AI. In addition, by analyzing trends in AI security and collaborating with research institutions and industries both in Japan and abroad, CREATE is building a sustainable and robust security infrastructure from a global perspective. Through these efforts, the center contributes to the development of next-generation security technologies and the realization of a safe and secure AI-native cyber society.

RESEARCH AREA

Security for AI

Ensuring the safety and reliability of AI is one of the key focus area. With the rapid adoption of large language models (LLMs) and other generative AI systems across society and industry, new risks such as adversarial attacks and unintended behaviors are emerging. CREATE focus on safety, security, and robustness to achieve resilient and trustworthy AI.

AI for Security

AI is used in cybersecurity to extract insights from vast amounts of security data. Real-time analysis enables the detection of unknown threats, malware, and emerging attack campaigns, enhancing the overall safety of cyberspace. In addition, CREATE addresses cybersecurity workforce shortages by advancing automated and autonomous defensive operations.

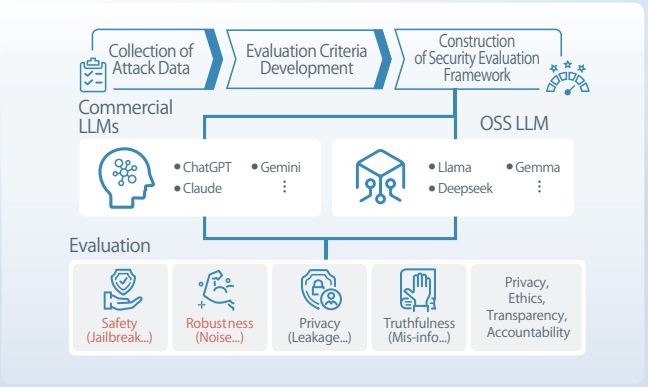
Trustworthy AI

To ensure the safe use of AI, it is important to make AI decisions understandable to humans. By developing explainable AI (XAI) mechanisms that enable intuitively understanding of the rationale behind these decisions, we aim to enhance trust in AI and to build a foundation for a society that can "understand and use" AI.

RESEARCH TOPICS

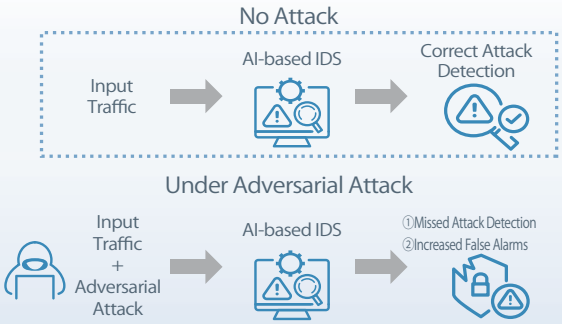
AI Security Evaluation Platform

This platform assesses the security risks of generative AI. Based on attack scenarios such as jailbreaks, harmful content generation, and information leakage, the platform objectively evaluates AI behavior using benchmarks. By comparing multiple AI models under the same conditions and visualizing the results, it enables intuitive understanding of differences in safety and risk tendencies. This supports safe use of AI and informed model selection.



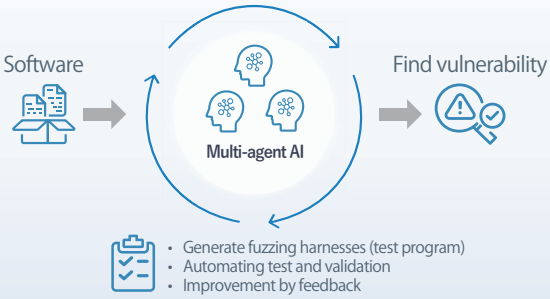
Adversarial Attacks against AI-based IDS

This research investigates how an attacker can mislead Intrusion Detection Systems (IDSs) that use AI to detect cyberattacks. These IDSs learn to tell the difference between normal and harmful network activity. However, attackers can make very small changes to the data that look normal to humans but confuse AI. As a result, the IDS might miss real attacks or incorrectly label safe activity as dangerous. Our goal is to understand these weaknesses and help make these IDSs more secure.



Automatic Fuzz Harness Generation by AI Multi-Agent System

We propose a multi-agent system in which multiple AI agents analyse source code, generate fuzz harnesses, and validate them automatically. Traditional fuzzing requires test code (fuzz harnesses), whose development depends on human effort and expertise. The system aims to accelerate the discovery of hidden bugs and security flaws. We also evaluate whether AI-generated harnesses can effectively uncover vulnerabilities.



OTHER RESEARCH TOPICS

- Automated SOC
- AI-enabled cyberattacks and defence
- Secure-by-design AI model
- Threat analysis of AI systems
- Explainability of AI
- AI and data privacy

CREATE
Center for Research on AI Security and Technology Evolution

National Institute of Information and Communications Technology
Cybersecurity research institute
Center for Research on AI Security and Technology Evolution

create-contact@ml.nict.go.jp https://create.nict.go.jp/en/

Tokyo Nihonbashi Office
Tokyo-Nihonbashi-Tower Bldg. 15F
2-7-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027

North-America Office (Washington DC)
NICT North-America Center
1020 19th Street NW Suite 880, Washington DC 20036