ITU Workshop on Generative AI: Challenges and Opportunities for Security and Privacy

Session 4: Panel discussion - Future directions

16:30 -- 17:30, 19 February 2024
Geneva, Switzerland

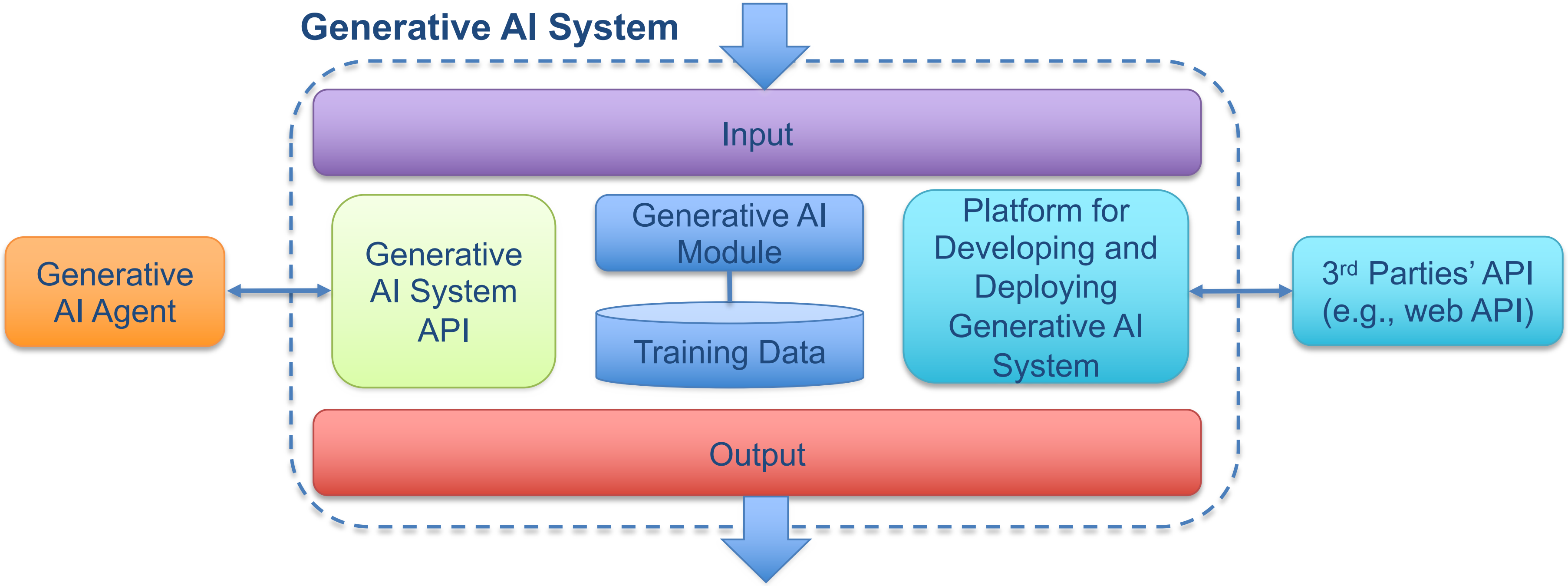
Session 4: Panel discussion – future directions

This session will explore opportunities for future standards and provide recommendations to ITU-T Study Group 17 for future work in this area.

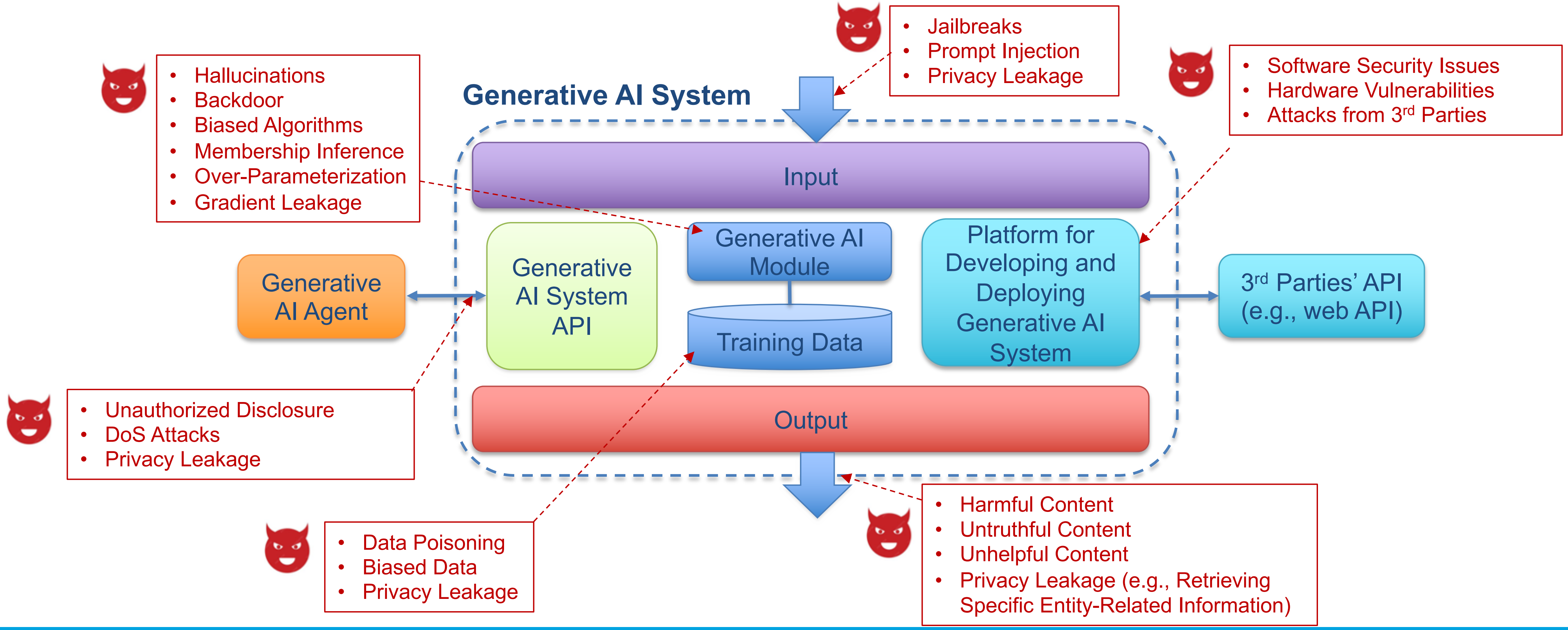
Members for the panel

- **Moderator: Zhiyuan Hu**, Vice-Chair, WP2/17, Question 2/17 Co-Rapporteur | Director, Security Research, vivo Mobile Communication Co. Ltd, China
- **Panelists:**
  - **Ann Cavoukian**, Executive Director, Global Privacy & Security by Design Centre
  - **Ramy Fathy**, Co-chair, ITU-T Focus Group on Artificial Intelligence and Internet of Things for Digital Agriculture" (FG-AI4A) | Director, Digital Services Planning and Risk Assessment, National Telecommunications Regulatory Authority of Egypt (NTRA), Egypt
  - **Clarisse Girot**, Head of the Data Governance and Privacy Unit, OECD
  - **Jabu Mtsweni**, Chief Researcher and Centre Manager for the Information and Cybersecurity Research Centre, Council for Scientific and Industrial Research (CSIR), South Africa
  - **Heung Youl Youm**, Chair, ITU-T Study Group 17, Security | Professor, Department of Information Security Engineering, Soonchunhyang University, Korea (Rep. of)

Overview of Generative AI System



Security and Privacy Problems



- Jailbreaks
- Prompt Injection
- Privacy Leakage

- Software Security Issues
- Hardware Vulnerabilities
- Attacks from 3rd Parties

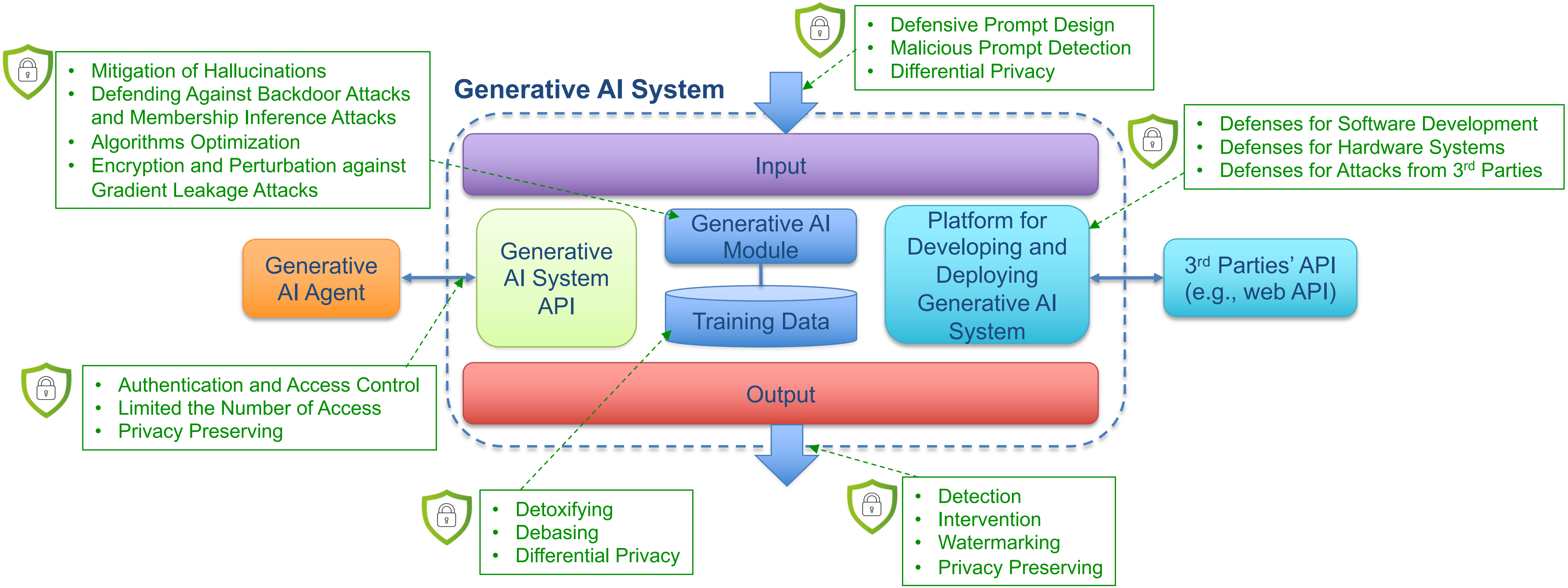
- Hallucinations
- Backdoor
- Biased Algorithms
- Membership Inference
- Over-Parameterization
- Gradient Leakage

- Unauthorized Disclosure
- DoS Attacks
- Privacy Leakage

- Data Poisoning
- Biased Data
- Privacy Leakage

- Harmful Content
- Untruthful Content
- Unhelpful Content
- Privacy Leakage (e.g., Retrieving Specific Entity-Related Information)

Security and Privacy Countermeasures



- Defensive Prompt Design
- Malicious Prompt Detection
- Differential Privacy

- Mitigation of Hallucinations
- Defending Against Backdoor Attacks and Membership Inference Attacks
- Algorithms Optimization
- Encryption and Perturbation against Gradient Leakage Attacks

- Defenses for Software Development
- Defenses for Hardware Systems
- Defenses for Attacks from 3rd Parties

- Authentication and Access Control
- Limited the Number of Access
- Privacy Preserving

- Detoxifying
- Debasing
- Differential Privacy

- Detection
- Intervention
- Watermarking
- Privacy Preserving

Generative AI for Security

- Cyber defense automation
- Threat intelligence
- Secure code generation and detection
- Identification of cyber attacks
- Malware detection
- Enhancing the effectiveness of cybersecurity technologies

Generative AI for Attack

- Social engineering attacks
- Phishing attacks
- Automated hacking
- Ransomware code generation
- Malware code generation

Questions for the panelists

- **Standardization Gaps:**
  a. What are the current gaps in international standards related to security and privacy of generative AI, generative AI for security, and how can they be addressed?
  b. How can future standards ensure interoperability and compatibility across different generative AI technologies and platforms?

- **Ethical considerations:**
  a. How can standards incorporate ethical guidelines to guide the development and use of generative AI technologies, for example, to prevent generative AI from being used for cyber attack?

- **Global Standards and Collaboration:**
  a. What ways can international collaboration and standardization efforts enhance the security and privacy of generative AI technologies?

- **Way forward for SG17:**
  a. Considering the discussion today, are there any topics from the 3 previous sessions that could lead to fruitful standardization work at SG17 with no duplication with other SDOs?
  b. What is your key takeaway for ITU-T SG17 based on your experience and discussion today?