

AICS 2026 Workshop Program
All times are SGT (Singapore Time)

08:55 – 09:00	Opening Remarks Remarks by Organizer
09:00 – 09:50	Keynote 1: Eugene Lim
09:50 – 10:30	Paper Session 1 • [15 min] Cybersecurity AI: Evaluating Agentic Cybersecurity in Attack/Defense CTFs. Francesco Balassone, Victor Mayoral-Vilches, Stefan Rass, Martin Pinzger, Gaetano Perrone, Simon Pietro Romano, Peter Schartner • [10 min] SceneGuard: Training-Time Voice Protection with Scene-Consistent Audible Background Noise. rui sang, Yuxuan Liu • [10 min] Training data membership inference via Gaussian process meta-modeling: a post-hoc analysis approach. Yongchao Huang, zhang pengfei, Shahzad Mumtaz
10:30 – 11:00	Break
11:00 – 12:00	Paper Session 2 • [15 min] Robustness of Operator Learning Methods with MLP-Based Kolmogorov-Arnold Networks. Dehami Kiryu, Atsushi Takabatake, Takaharu Yaguchi • [15 min] Jailbreaking the Security of LLMs: How Targeted Nullspace Perturbations Unmask Hidden Vulnerabilities. Vishal Pramanik, Maisha Maliha, Susmit Jha, Sumit Kumar Jha • [15 min] Improving Neural Network Robustness to Convolutional Perturbations Through Certified Training. Benedikt Brückner, Alessio Lomuscio • [20 min] AI-based Traffic Modeling for Network Security and Privacy: Challenges Ahead. Dinil Mon Divakaran
12:00 – 01:30	Lunch (on your own)

1:30 – 2:20

Keynote 2: SUN Jun

2:20 – 3:30

Paper Session 3

- [15 min] SPUN: Spectral Projection–based UNlearning in Hyperdimensional Computing. Vishal Pramanik, Olivera Kotevska, Alvaro Velasquez, Susmit Jha, Sumit Kumar Jha
- [15 min] An open framework for AI Security. Víctor Mayoral-Vilches, Luis Javier Navarrete Lozano, María Sanz-Gómez, Lidia Espejo Salas, Martiño Crespo Álvarez, Francisco Oca González, Francesco Balassone, Alfonso Glera-Picón, Unai Ayucar Carbajo, Jon Ander Ruiz-Alcalde, Stefan Rass, Martin Pinzger, Endika Gil-Uriarte
- [15 min] Large Language Model-Based Reward Design for Deep Reinforcement Learning-Driven Autonomous Cyber Defense. Sayak Mukherjee, Samrat Chatterjee, Emilie Purvine, Ted Fujimoto, Tegan Emerson
- [15 min] Lost in Translation? A Comparative Study on the Cross-Lingual Transfer of Composite Harms. Vaibhav Shukla, Hardik Sharma, Adith N Reganti, Soham Wasmatkar, Bagesh Kumar, Vrijendra Singh
- [10 min] SubstratumGraphEnv: Reinforcement Learning Environment (RLE) for Modeling System Attack Paths. Bahirah Adewunmi, Edward Raff, Sanjay Purushotham

3:30 – 3:50

Break

3:50 – 5:00

Paper Session 3

- [15 min] Fact or Hallucination? An Entropy-Based Framework for Attention-Wise Usable Information in LLMs. Vishal Pramanik, Alvaro Velasquez, Susmit Jha, Sumit Kumar Jha
- [15 min] Peering Behind the Shield: Guardrail Identification in Large Language Models. Ziqing Yang, Yixin Wu, Rui Wen, Michael Backes, Yang Zhang
- [10 min] Beyond BeautifulSoup: Benchmarking LLM-Powered Web Scraping for Everyday Users. Arth Bhardwaj, Nirav Diwan, Gang Wang

- [10 min] Orchestrating Cybersecurity Tools: An AI Agent for Automated Malware Triage and Attribution. Larry Milner
- [10 min] TopoReformer: Mitigating Adversarial Attacks Using Topological Purification in OCR Models. Bhagyesh Kumar, A S Aravindhakshan, Akshat Satyanarayan, Ishaan Gakhar, Ujjwal Verma
- [10 min] Cybersecurity AI: Humanoid Robots as Attack Vectors and How to Defend Them. Víctor Mayoral-Vilches, Andreas Makris, Kevin Finisterre

05:00 – 5:05

Closing Remarks / Adjourn