CECS
CENTER FOR EMBEDDED & CYBER-PHYSICAL SYSTEMS
UNIVERSITY OF CALIFORNIA · IRVINE

CECS Seminar



"Byzantine Agents: From Untrustworthy Models to Provably Robust Inference"

Professor Ihsen Alouani

Associate Professor at the Centre for Secure Information Technologies (CSIT), Queen's University Belfast, and a Faculty Affiliate of the Hardware AI Governance Lab at the Universitv of Oxford

Wednesday, December 9th
1:00-2:00 p.m.
Location: EH 2430

Abstract:

Classical security assumes that instructions can be kept separate from data and that correct behaviour can be specified. Language models offer neither, and the consequences compound as we move from a single model to the systems built around it. At the model level, confidentiality fails from the inside. We show that training-set membership can be inferred from a model's self-attention patterns, and we prove that a model can be forced to memorise specific information that the attacker does not know in advance. Integrity is equally fragile: by manipulating where a model attends, an adversary can redirect what it does. One layer up, a common mitigation is to wrap the model in learned detectors. We show that evasive injections bypass them, because a learned detector inherits the adversarial weaknesses of the model it guards. Since individual models can be neither isolated nor certified, I argue that trust must move from the component to the composition, and present provably robust federated inference for LLM agents, which gives guarantees on the aggregated output when a bounded fraction of participating agents are adversarial.

Biography:

Ihsen Alouani is an Associate Professor at the Centre for Secure Information Technologies (CSIT), Queen's University Belfast, and a Faculty Affiliate of the Hardware AI Governance Lab at the University of Oxford. His research addresses the security and privacy of AI systems across the computing stack. He has published over 90 peer-reviewed papers, at venues including CVPR, EMNLP, ASPLOS, PoPETs and IEEE TIFS/TCAD. He leads research funded by UKRI, Horizon Europe, the UK's Laboratory for AI Security Research (LASR), and a US-Ireland partnership with NSF and UC Riverside. He is an IEEE Senior Member, PC member of several flagship venues, and served as AI/ML Security and Privacy Track Chair at DAC 2023 - 2025.

Hosted By: Prof. Mohammad Al Faruque