For the call: HORIZON-CL3-2026-02-CS-ECCC-02 — SecureAI (Enhancing the Security, Privacy and Robustness of AI Models and Systems)

Horizon Europe, Cluster 3 | Deadline: 15 September 2026

SecureGenAI: Adversarial Resilience for Generative AI Systems

2026-04-23

Jianguo Ding (dji@bth.se), Kurt Tutschku (ktt@bth.se),
Roman-Valentyn Tkachuk (rvt@bthse,)

Department of Computer Science – Secure and Distributed Systems
Blekinge Institute of Technology (BTH)
Karlskrona, Sweden

Member of CYBERCAMPUS SVERIGE

PROJECT IDEA & OBJECTIVES

Challenges: GenAI models (LLMs, GANs, diffusion) are increasingly used in cybersecurity and critical infrastructure — yet remain **vulnerable to adversarial manipulation, data poisoning, deepfakes, and privacy leakage**, creating new attack surfaces. These defences must also work in distributed, privacy-sensitive settings, ensuring trustworthiness and EU AI Act compliance.

Objectives:

- **Robust GenAI Defence:** Adversarial training, attack detection, and recovery for GANs, LLMs, and diffusion models
- **Deepfake & Misuse Detection:** Deployment-ready benchmarks and detection pipelines for synthetic media and manipulated data
- **Trustworthy & Distributed GenAI Deployment:** Privacy-preserving, auditable GenAI via federated learning, blockchain-verified provenance, and TEEs (Trusted Execution Environment) — ensuring EU AI Act compliance for government and enterprise deployments

Expected Outcomes & Impacts

- **Adversarial robustness toolkit for GenAI**
  Attack detection and recovery for LLMs, GANs, and diffusion models — building on our adversarial defence methods
- **Deepfake & synthetic content detection pipeline**
  Benchmarks for AI-generated media, building on our deepfake detection work (IEEE TITS, IET ITS)
- **Trustworthy GenAI deployment framework**
  Blockchain-anchored federated training with secure aggregation and TEEs for trusted in-house GenAI deployment
- **Critical infrastructure pilot validation**
  2–3 pilots: smart grids, autonomous transport, government decision support

BTH Expertise & Desired Partners

BTH Track Record:

- Research: GenAI for Cybersecurity, Adversarial ML, Federated Learning, Trusted Distributed Systems, Digital Sovereignty
- European Projects (cont. since 2017): H2020 Bonseyes, H2020 BonsApps, HE dAIEDGE, Celtic+ CISSAN
- International network / potential partners: Switzerland, Finland, Spain, Germany, Turkey, Hungary
- National networks: Member of **Cybercampus,** Swedish Industrial Graduate School on Cybersecurity

Member of CYBERCAMPUS SVERIGE

Proposed role for BTH: WP lead – e.g. GenAI robustness & adversarial defence and/or trustworthy distributed AI deployment

Desired Partners:

- **Research institutes:** adversarial ML, LLM security, deepfake detection, privacy-enhancing technologies
- **Industry / SMEs** — GenAI platforms, AI cybersecurity products, CI operators (energy, transport, gov IT)
- **Government / CERTs** — pilots for GenAI decision support, AI Act compliance, deepfake response, governance

Contact: dji@bth.se