New Security Controls for Decentralised Agentic AI

Supervision team
Main Supervisor: Shishir.Nagaraja@newcastle.ac.uk

Research project
We are rapidly approaching a world saturated with powerful, decentralised AI agents. These agents and their underlying models have astonishing capabilities, but our ability to actually control them—to dictate how, where, and by whom they are used—is almost non-existent. For Edge AI in particular, trust must become a mechanical property, engineered into the system from the ground up to withstand both internal and external adversaries. The goal of this project is to build the cryptographic and architectural primitives that make this possible.

The project aims to develop AI models that do more than process data: they must be able to negotiate with each other under adversarial conditions, using a new language of interaction grounded in post-quantum cryptographic proofs. A first strand of work will focus on zero-knowledge AI handshakes, where models exchange data and capabilities only after mutually verifying each other's credentials and intentions through zero-knowledge proofs (ZKPs). This allows them to prove entitlement to knowledge without revealing the secret itself, preventing unauthorised access and limiting information leakage. A second strand will develop threshold classifiers, where no single entity holds unilateral control. Instead, classification capability requires consensus—for example, three out of five nodes must cryptographically sign and unlock an action. This reduces the risk that a single compromised agent could trigger malicious outcomes.

Further research will explore access-controlled Edge AI, in which model functionality is not fixed but unlocked dynamically via authorisation keys. This allows a single deployed model to vary its capabilities instantly based on the credentials of the user, providing a direct mechanism for privilege management, rapid revocation, and containment of compromised entities. Alongside this, the project will embed end-to-end auditable workflows, incorporating verifiable confidentiality and tamper-evident audit trails directly into the AI's operation. Every decision and interaction will generate a forensic record, providing both accountability and a mechanism for assigning liability when failures occur. Finally, the project will investigate deniable AI models, which deliberately obscure the provenance of training data. These "fogged" models protect privacy and act as a defensive measure against model inversion and membership inference attacks, ensuring adversaries cannot prove what data the model was trained on.

This research moves beyond building smarter AI to building civilised AI: systems whose power is tempered by verifiable constraints, whose actions can be audited, and whose failures can be contained and attributed.

This project is supported by National Edge AI Hub https://edgeaihub.co.uk/ and is open for self-funded students.