

Explainable and Trustworthy AI in Cybersecurity

Fully Funded PhD Studentship

School of Internet of Things | Xi'an Jiaotong–Liverpool University

Scholarship support

- Full tuition fee waiver.
- Living stipend of RMB 5,000 per month for 36 months, totalling RMB 180,000.
- Academic conference funding of RMB 16,500 in total for the entire PhD. This is not an annual allowance.

We invite applications from motivated candidates seeking doctoral research in this area. Scholarship administration and eligible conference expenditure are governed by the formal award terms and University procedures. The start date will be agreed in line with the relevant doctoral admissions intake.

Project overview

When an AI system flags a cyber threat, can an analyst understand the evidence, test the reasoning and recognize when the prediction should not be trusted? This PhD will investigate neuro-symbolic methods that combine learned representations with explicit security knowledge. The candidate will develop and evaluate models for a focused cybersecurity task, examining predictive quality, explanation faithfulness and robustness under changing or adversarial conditions. The aim is to support accountable security decisions, with uncertainty and limitations made visible to the analyst.

Research background and rationale

AI-based intrusion detection and security analytics can identify patterns in large volumes of telemetry, but predictive performance alone does not establish whether a decision is dependable. An explanation may appear plausible without faithfully representing the model's decision process. A model may also become unreliable when traffic changes, evidence is incomplete or an attacker manipulates its inputs.

Neuro-symbolic methods offer a way to connect statistical learning with explicit rules, relationships or constraints. In this project, symbolic knowledge will be treated as a testable component rather than an automatic source of correctness: rules may conflict, become outdated or fail to cover an unfamiliar attack. The research will examine how these limitations should affect both predictions and explanations.

The scientific focus is whether a defined combination of learning and reasoning improves the quality of security decisions relative to neural, rule-based and post-hoc explanation baselines. A focused task, initially intrusion detection or alert triage, will provide a tractable basis for experimentation before any extension to distributed AIoT settings.

Research objectives and questions

The candidate will refine the research questions with the supervisory team following a critical literature review and initial feasibility experiments. The proposed questions are:

- How can security knowledge be integrated with learned representations to produce explanations that are faithful to the decision process and linked to observable evidence?
- How should a neuro-symbolic model handle conflicting rules, missing evidence and distribution shift, including when to abstain or request analyst review?
- Under what conditions does the approach improve robustness or analyst decision quality, and what predictive and computational trade-offs does it introduce?

Research methods and evaluation

Task definition and defensible baselines. Select a bounded detection or triage task, identify the intended analyst decision and define an attacker model. Establish neural, interpretable rule-based and suitable post-hoc explanation baselines. Use licensed or approved datasets with documented preprocessing, label limitations and train/test separation.

Neuro symbolic model development. Investigate an appropriate representation of security knowledge, such as rules, a knowledge graph or logical constraints. Design an interface between learned representations and reasoning, with traceable evidence for predictions. Include mechanisms for uncertainty, rule conflict and insufficient evidence where required by the task.

Explanation and robustness evaluation. Measure predictive performance with metrics suited to class imbalance, including precision–recall measures and false-alarm rates. Test explanation faithfulness through controlled perturbation or counterfactual checks, and assess stability and consistency. Evaluate defined adversarial manipulations, missing evidence and temporal or cross-dataset shift without leaking test data into model development.

Decision support and reproducibility. Compare the full method with ablations that remove or alter its reasoning components. Measure calibration, abstention behaviour, latency and resource cost. Assess analyst usefulness through a predefined task-based protocol where participant access and ethics approval permit; otherwise use expert-reviewed cases without claiming validated human benefit. Package experiments, configurations and limitations for independent reproduction.

Expected doctoral contribution

The intended doctoral contribution is an evaluated neuro-symbolic method for evidence-linked cybersecurity decisions, with a clear account of when its explanations are faithful and when the model should defer. Expected research outputs include a prototype, benchmark comparisons and an evaluation protocol. Explainability will not be presented as a guarantee of robustness, and the thesis will not attempt to build a complete enterprise security operations platform.

Research training and environment

Close collaboration with our industry partner will be a central part of this PhD. The candidate will work with partner specialists to define cybersecurity tasks, establish baselines and validate models within application workflows. Experiments will examine predictive performance, explanation faithfulness, robustness, analyst usefulness and system integration. Partner feedback will inform model refinement and analysis of the results.

Partner data and system access will follow agreed confidentiality, intellectual property, ethics and security arrangements. Approved benchmark and sandbox experiments will complement industry validation and support reproducibility.

Supervisory review will guide literature review, baseline experiments, method development and thesis evaluation, maintaining a focused, original doctoral contribution alongside industry collaboration.

Candidate profile

- A strong academic background in computer science, artificial intelligence, cybersecurity, mathematics, electronic engineering or a closely related field, subject to the University's doctoral admission requirements.
- Sound foundations in machine learning, probability and experimental evaluation, with practical Python programming and experience using a framework such as PyTorch.
- Evidence of research potential through a dissertation, substantial ML/security project or equivalent work, and the ability to critically examine datasets, baselines and model claims.
- Ability to read research literature, write clearly in English and work independently with academic guidance.

Desirable experience

Experience with knowledge representation, logical reasoning, explainable AI, adversarial machine learning, intrusion detection or security analytics is advantageous. Applicants may bring strength in either AI methods or cybersecurity and develop the complementary expertise during the project.

Application and academic enquiries

Academic contact: Professor Dijiang Huang, School of Internet of Things, Xi'an Jiaotong–Liverpool University.

Email: Dijiang.huang@xjtlu.edu.cn

For an initial academic enquiry, please provide:

- A curriculum vitae describing your academic background and relevant research or technical experience.
- Academic transcripts and available degree information.
- A short statement of research interests explaining your fit with this project and a question you would like to investigate.
- Where available, a dissertation abstract, publication, technical report or repository showing your own contribution. Do not disclose confidential material.
- Your preferred start date and current study or employment commitments.

Suggested email subject: PhD Application — Explainable and Trustworthy AI — Your Name

An academic enquiry does not replace the formal doctoral application. Admission and the scholarship award are subject to the University's formal selection and offer process. Detailed eligibility and application documentation should be confirmed through that process.

Research areas

Neuro-symbolic AI; explainable AI; cybersecurity analytics; adversarial robustness; uncertainty; AIoT security