The Alan Turing Institute

AI for Cyber Defence for Research Centre – Machine Learning for Security and Privacy

TIN-ATI-002

About the Organisation

AICD is a research centre at the Alan Turing Institute that is aiming to transform the science of computer security and privacy using machine learning (ML) and autonomous decision making. We believe that ML, in particular deep reinforcement learning (DRL) and transformer models, has a significantly underdeveloped potential to autonomously defend, attack and quantify the security of computer networks and systems. Our mission is to develop the fundamental and applied techniques in ML that will allow us to demonstrate this potential.

Role Description and Responsibilities

Background

DRL has demonstrated superhuman performance in a range of games including chess, Go and Dota. DRL has also made an invaluable contribution to the latest foundational models such as GPT and Bard, helping to tune their outputs to human preferences. AICD conducts fundamental and applied research on the application of ML to computer security and privacy challenges, with the aim too of establishing superhuman levels of performance.

Project: Investigating the use of ML/DRL to autonomously tackle computer security and privacy.

The goal of this research is to investigate the application of ML/DRL to autonomously tackle computer security and privacy. The research is expected to be significantly experimental and includes the following objectives:

- To capture the relevant existing literature on the use of autonomous methods for security and privacy in a computer networks and system.
- To identify potential autonomous techniques (or technique enhancements) that could be used to advance a security or privacy challenge.
- To evaluate the proposed methods on a simulated, emulated, or real environment as justified by the security and privacy context of the research.

Methodology

The research will be conducted in several stages including:

1. **Full challenge specification.** Computer security and privacy encompasses a broad range of discrete challenges. A specific security and privacy direction will be chosen

by each researcher based on their previous expertise and a review of the literature. If no direction is identified as preferrable then one will be chosen from autonomously identifying a systems security flaw, autonomously defending an enterprise network through self-play and autonomously defeating a capture-the-flag (CTF) challenge.

2. **Literature review.** A more thorough review of the literature will be conducted to identify all related work. This will include any autonomous or computer-assisted methods that might not qualify as DRL but which nonetheless provide baselines for comparison.
3. **Technique proposal.** A specific set of techniques, architecture and methodology for tackling the chosen challenge will be developed by the researcher with guidance from the centre leads. This will include identifying a suitable environment for training and evaluating autonomous agents.
4. **Experimental research.** The proposed methods will be executed, and the results will be recorded. Where time permits, an iterative approach to method refinement and execution will be performed to improve the results where possible.
5. **Impact analysis and write up.** A written report or academic article will describe the potential of DRL, and any associated methods identified during the research, to autonomously tackle computer security and privacy challenges.

The two researchers appointed to this role will conduct research with AICD colleagues towards applied advances in autonomous computer security and privacy. The researchers will be responsible for:

- Conducting research and framing a security and privacy challenge in the context of autonomous decision-making capabilities.
- Developing and testing autonomous algorithms.
- Communicating findings and working with colleagues to optimise their results.

Some of the ML techniques likely to be of relevance include large language models, transformers and attention, multi-agent reinforcement learning and DRL.

Expected Outcomes

The expected outcomes are as follows:

- A report or academic paper describing the research findings identified during the internship.
- A final presentation to stakeholders.

Supervision and Mentorship

The role will be supervised by Chris Hicks and Vasilis Mavroudis who co-lead AICD however the wider team will also be available for collaboration, supervision and mentorship.

Person Specification

The ideal intern is passionate about exploring the intersection between computer security and ML. Ideally:

- You have found a vulnerability (known or novel) in a computer system, network, or algorithm before and maybe even written up the results;

- You have worked with frameworks for ML such as pytorch, tensorflow, jax, etc and might even be familiar with gym, rllib and huggingface;
- Problem-solving and analytical skills;
- Strong written and verbal communication skills;
- Ability to work independently and as part of a team;
- Excellent collaboration skills, with the ability to work effectively in a team environment.

Internship Logistics

Salary: £40,000 per annum pro-rata

Duration: 3-6 months full time. We are also open to part-time roles for 6-12 months.

Location: London. Flexible/hybrid working (including fully remote) is possible.

A background check will need to be completed before the successful candidate(s) can be onboarded.

This position is available for 2 interns.