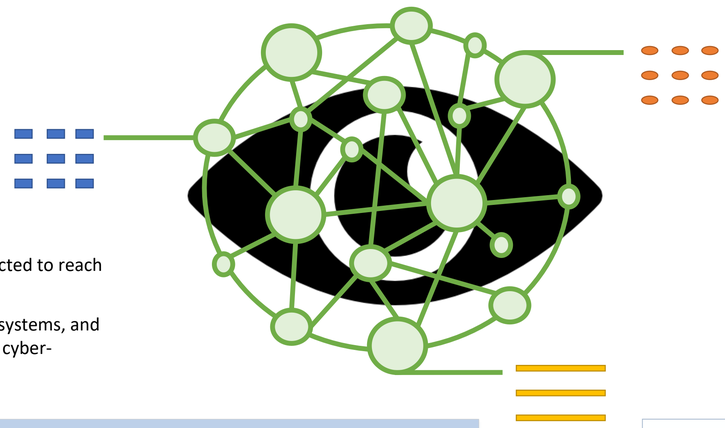
CAMLIS
October 2023

Can We Trust Autonomous Cyber Defence for Military Systems?

Autonomous Resilient Cyber Defence (ARCD) Programme
High-Risk and Disruptive Options (HRDO) Project

INTRODUCTION



An overview of the cyber defence problem

In 10-15 years' time, machine speed cyber-attacks in the defence domain are expected to reach a pace and volume beyond that which can be managed by humans alone.

In addition to this, due to both the increasing complexity of military networks and systems, and the sophisticated approaches of aggressors, it is becoming increasingly difficult for cyber-defenders to respond quickly and effectively to incidents.

What is the ARCD programme?

The Autonomous Resilient Cyber Defence (ARCD) programme, led by the UK Defence Science and Technology Laboratory (Dstl) is undertaking research to "exploit advances in Information Technology (IT), Operational Technology (OT) and Machine Learning (ML) to develop self-defending, self-recovering concepts for military operational platforms and technologies that provide a substantial improvement in cyber resilience."

To date, over 60 research tasks have been exploring State Of The Art (SOTA) cyber defence concepts, bringing together expertise across ML, and cyber security. The programme objectives are to achieve the following:

- The creation of a concept demonstrator capable of autonomously responding to cyber-attacks in the context of a military environment and mission.
- Enhance cyber and ML skills in the UK supply chain.
- Further developing understanding of strengths and limitations of ML technologies and their application into Cyber Defence.

Why is eXplainable AI so important?

Explainability is key to developing a user trust, and a particularly crucial element when considering the recent progress on, and therefore challenges arising from, ML methods such as Reinforcement Learning (E.Puiutta et al, 2020) and the ARCD context (automating responses in a fast-paced, safety critical environment). Trust is built, inter alia, through several layers of explanation, such as:

- accurate representation of system states,
- communication of the operational limits where confidence is reduced,
- evidencing the sensemaking process of expected and unexpected results,
- explanations of context aware reasoning and decisions taken.

Finally, "how best to provide meaningful and fine-tuned explanations to the different stakeholders of the current or projected decisions" underpins each of these aspects in enabling effective and efficient user comprehension during deployment, or through providing accurate evidence and explanations during validation, verification, and assurance processes.

... And the HRDO project?

ARCD has embraced a high-risk high-reward appetite to drive success in the development of Generation After Next (GAN) cyber defence concepts within the HRDO project. 'Success' here is defined as pushing technical boundaries, exploring challenges and learning lessons from technical failures to improve understanding of the strengths and limitations of ML technologies and their application to cyber defence. To date, research has focussed on the following problem areas:

Causal Inference · Pattern of Life · eXplainable AI · Knowledge Graphs · Quantum Machine Learning · Generalisability · Data Efficient Decision-Making

The HRDO research teams have employed SOTA techniques including RL; Deep Learning (DL); Supervised, Semi-Supervised and Unsupervised approaches; Genetic Algorithms (GAs); Transformers and Large Language Models (LLMs); and Quantum Machine Learning (QML).

WHAT DOES THE RESEARCH SAY?

XAI and Policy Dissection

[Task Complete. Delivered by Cambridge Consultants]

The question: Whether policy dissection (Q.Li et al, 2023) can be integrated within cyber environments to analyse neural network activity, thereby identifying patterns of behaviour?

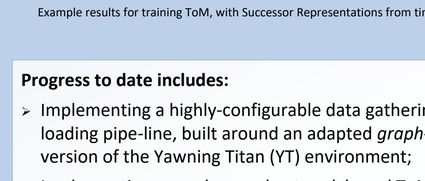


Figure 1: Actions on nodes over a full episode – all actions on node 2 (left) and all 'repair' actions on node 8 (right)

Figure 2: Correlation of two neurons to single observations

The answer: this work shows potential to address black-box concerns. The team have successfully shown:

- Extraction of the neuron activations throughout the network to facilitate explainability within a cyber security environment;
- That the importance of different neurons of the Neural Network (NN) can be extracted for different actions taken by the RL;
- How the extracted information can be used to determine the state of the cyber environment as 'seen' by the RL;
- The potential of using additional NNs to address the challenges in connecting the environment state to the resulting actions using the extracted activations in the final hidden layer of the network.

The challenges of note:

- Latter neurons do not encode all possible observations in the cyber environments, discarding 'unimportant' information, so does the RL encode everything it needs?

The recommendations include enhancing development of the 'Juvenal' network to provide real-time feedback alongside natural language descriptions.

Additional context: this technique originally aligns the intermediate representation of the learned neural controller with the kinematic attributes of the agent behaviour (so that stimulation of certain neurons can produce a desired action).

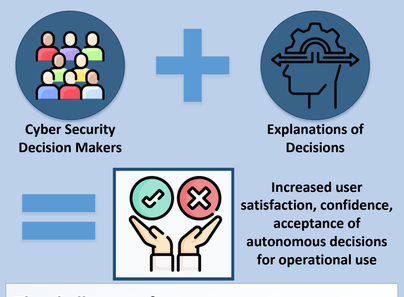
eXplainable RL (XRL)

[Reaching completion. Delivered by BMT with Mind Foundry]

The question: How to develop and train an eXplainable RL (XRL) agent, where explanations fall under three different categories: Feature Importance; Learning process and Markov Decision Process and Policy-Level; and, how to best visualise the explanations made by the RL agent?

The answer: general-purpose XRL techniques enhance the understanding of defensive actions in multi-agent RL (MARL) cyber defence environments and can be provided to cyber analysists for real investigations in a GUI.

- Explanations should always be developed with the end user in mind;
- Postmortem and Step-by-step explanations are different and as such require different design considerations;
- The 'Why' to agent actions can be provided by Scenario replay, Shapley values and reward function decomposition methods;
- Anecdotal and user-based evaluations show "Context and Remedial Actions" visual explanations are required to support the explainable RL elements;
- XRL Explanations are useful for debugging/ training during Algorithm Development.



The challenges of note:

- During evaluation processes it is important to disambiguate the quality of agent performance from the quality of an explanation.

The recommendations research into XRL targeting higher fidelity environments to close the Sim to Real gap.

Additional context: This research team split out the 'what' (happened and has been done), 'when', 'where', 'why' (have these actions been taken) and 'how', in order to determine how the XRL approaches apply to, and best present, the various elements of explanation required by cyber analysts.

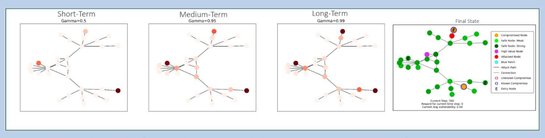
The results of their initial literature review (S.Milani et al, 2022) consider how a problem should split along the algorithmic and usability lines to bridge the gap between abstract mathematical outputs (targeting ML researchers) and end-user interfaces demonstrating typically bespoke visualisations (targeting cyber analysts) in the form of a 'dashboard'.

The user is able to gain a sense of scenario history and agent actions on nodal diagram as the visualisations have been grouped into in a step-by-step and postmortem views.

Theory of Mind (ToM)

[Task ongoing. Delivered by BAE Digital Intelligence]

The question: Whether XRL methods such as Machine Theory of Mind (ToM) (N.C. Rabinowitz et al, 2018) approaches allow a blue agent to learn so it can make predictions on the beliefs, goals and desires of red RL agents?



Example results for training ToM, with Successor Representations from time step 0

Progress to date includes:

- Implementing a highly-configurable data gathering and loading pipe-line, built around an adapted *graph-based* version of the Yawning Titan (YT) environment;
- Implementing a graph neural network based ToMnet through which to apply their graph based YT environment;
- Evaluating the graph-based ToM network against a cyber-defence scenario designed and called the *Hot-Desking User* problem. ToMnet is tasked with predicting:
  - ✓ which user the red agent is going to be targeting in the current episode,
  - ✓ the exact node, and
  - ✓ what the user's likely path will be.

The answer: the research evaluates the extent to which a graph neural network based ToM model can help humans better understand the decisions made by cyber-attacking agents through predict their long term goals and likely trajectory through the network.

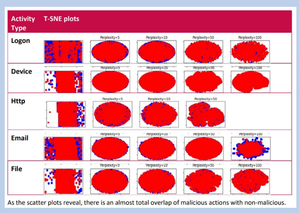
Preliminary results include:

- ToM finds it easier to predict the user that the red agent is targeting compared to the exact node. However, the confusion matrix show that ToM networks generally predict the correct branch;
  - ✓ node predictions become more challenging for larger topologies as the number of nodes increases.
- Predictions with respect to the targeted users are better for larger (90 node) networks compared to smaller ones (30 nodes);
  - ✓ The team hypothesize that the smaller network is more challenging due to the blue agent often winning games, while it struggles in the larger network.
- t-SNE visualisations of embeddings obtained from the ToM network's character network show clusters emerging that align with the agent's preference over users.
  - ✓ Early observations show non-overlapping clusters which often also coincide with distinct branches of the network frequented by the targeted user.

A Neuro-Symbolic AI (NeSy) Approach

[Task complete. Delivered by Decision Lab]

The question: How might a white-box ML model be developed to provide interpretable and understandable explanations of results in the cybersecurity domain and how does it compare to black-box AI models?



The answer: A Logic Tensor Network (LTN) model, combining the high predictive power of NNs with the inherent explainability of symbolic AI, was successfully integrated with a user interface (UI) to meet the XAI objectives. However the model exhibited poor performance when classifying malicious insider behaviour.

- The LTN model includes the functionality to interact and query the model showing clear and effective application into an XAI tool;
- Poor model outcomes hinder the ability to show the benefits that this XAI approach could bring;
- This poor performance is in part due to the overlap between malicious and non-malicious users;
- The UI design produced a clear and user-friendly method for a user to access and interact with the model on a case-by-case basis as well as querying individual predicates. A series of displays and dashboards allowed the user to understand the different levels of risk faced by their organisation to malicious insiders.

Additional context: This shows how lessons learned can be as powerful as classically successful results.

XAI as an Attack Vector

[Task ongoing. Delivered by Intellium AI]

Image generated by Intellium AI when prompting Generative AI with "Autonomous Resilient Cyber Defence"

The question: How might an inversion attacker attempt to reconstruct training data using model explanations as an additional attack vector and how might we apply game-theory and privacy-enhancing approaches to develop autonomous response and recovery mechanisms for XAI models that work on network data?

To date the team have:

- Developed and tested three novel mechanisms that help an AI system recover from an inversion attack: Feature Shuffling, Feature Cipher Shifting, Adding Noisy Features;
- Provisionally found that applying the combination of Feature Cipher Shifting and Adding Noisy Features mechanisms results the optimal results: an average increase of 32.96% in the data reconstruction error of the inversion attack.

DO YOU TRUST THE AI YET?

HRDO's high-risk high-reward approach and the 'successes' presented here will inspire further research into the application of novel XAI techniques to cyber defence, to build trust in autonomous agents and enable them to be confidently deployed and used on real networks.

The XAI focussed research tasks within the HRDO project can be seen to:

Successfully apply SOTA technologies to cyber defence environments,

Develop algorithms which can leverage information from within different parts of black and white-box models, while moving away from limited rules-based approaches, and

Apply these techniques to generate a broad range of explanations required by cyber defence stakeholders, covering:

'What does the agent understand about its environment?'

'What does the agent understand about the factors contributing to suggested response/decisions?'

'What does the agent understand about the question being asked?'

Advancing the techniques and tools with which stakeholders may access explainability layers will also benefit the remaining ARCD research areas.

This enables researchers and developers to demonstrate intelligent, sensible behaviour or alternatively to probe more deeply into the erroneous areas of understanding explained by the XAI agents.

Finally, XAI can support identification of novel cyber decision strategies, where a feature may be seen that is neither typical of a user's expectation nor evidently incorrect.

References

E.Puiutta et al. (2020). Explainable Reinforcement Learning: A Survey. https://researchgate.net/publication/343751190_Explainable_Reinforcement_Learning_A_Survey.

N.C.Rabinowitz et al. (2018). Machine Theory of Mind. http://proceedings.mlr.press/v80/rabinowitz18a/rabinowitz18a.pdf.

Q.Li et al. (2023). Human-AI Shared Control via Policy Dissection. https://arxiv.org/pdf/2206.00152.pdf.

S.Milani et al. (2022). A Survey of Explainable Reinforcement Learning. https://arxiv.org/pdf/2202.08434.pdf

Acknowledgements

The research (ARCD) funded by Frazer-Nash Consultancy Ltd. on behalf of Dstl, an executive agency of the UK Ministry of Defence providing world class expertise and delivering cutting-edge science and technology for the benefit of the nation and allies.

ARCD Track 1: ARCD Concepts
Email: arcd@fnc.co.uk
Author: Alex Revell