Engineering and Physical Sciences Research Council

University of Exeter

Engineering and Physical Sciences Research Council Doctoral Landscape Award

PROJECT TITLE: Understanding emergent behaviour in multi-agent reinforcement learning

Lead Supervisor: Xiaoyang Wang

Co-Supervisors: Chunbo Luo

Webpage: https://experts.exeter.ac.uk/40473-xiaoyang-wang

Project details: Multi-agent systems have demonstrated success in complex coordination tasks such as games, resource allocation, and collaborative robotics such as drone swarms. Despite their effectiveness, understanding emergent behaviours in these systems remains a challenge, mainly because policies are represented using deep neural network architectures. To enhance interpretability, recent work has shifted focus from traditional measures of performance (reward, evaluations against human experts, etc) to better understanding emergent behaviours in multi-agent systems. Existing research such as post-hoc interpretability techniques are limited in their ability to provide actionable insights. Intrinsically interpretable models, on the other hand, offer a promising solution by forcing decision-making to pass through a human-understandable latent space, enabling both behavioural analysis and causal interventions. This project focuses on emergent behaviours. Specifically, we aim to develop intrinsically interpretable models that improve our understanding of how agents coordinate and adapt in complex environments. Potential approaches include concept-based approaches, such as Concept Bottleneck Policies (CBPs), to enforce decision-making processes that rely on human-understandable representations. The goal is to develop approaches that enable real-time behavioural analysis and support test-time interventions to diagnose coordination successes and failures. The outcomes of this research will contribute to safer and more reliable AI systems by making multi-agent behaviours understandable and predictable. Additionally, the findings will support applications in domains such as healthcare, autonomous vehicles, and drone swarms, where interpretability is critical for trust and accountability.

Project specific requirements: Familiar with machine learning and deep learning.

Potential PhD programme of study: PhD in Computer Science

Department: Computer Science

Location: Innovation Centre, Streatham Campus

Please direct project specific enquiries to: Dr Xiaoyang Wang, x.wang7@exeter.ac.uk

Please ensure you read the entry requirements of programme to which you are applying.

To apply for this project please click here.