EPSRC DTP PhD Research Project

Project Title: Robustness Evaluation in Reinforcement Learning

Primary Supervisor details: Dr. Ronghui Mu**,** **r.mu2@exeter.ac.uk**, **https://scholar.google.com/citations?user=td_ct8MAAAAJ&hl=en**

Additional Supervisory team details: Xiaoyang Wang**,** **x.wang7@exeter.ac.uk** **https://wang-xiaoyang.github.io/**

Department: Computer Science

Location: Innovation Centre

PhD Programme: PhD in computer science

Project Description:
This project aims to explore the challenges and opportunities in enhancing the reliability, stability, and generalizability of reinforcement learning (RL) algorithms. We will first investigate the inherent vulnerabilities of RL algorithms, such as their sensitivity to environmental perturbations, adversarial attacks, and distributional shifts. Alternatively, we may analyze the impact of noisy, incomplete, or misleading feedback on the learning process and decision-making accuracy of RL agents, which can be considered in the training processes of large language models (LLMs).

The successful candidate is also encouraged to delve into the theoretical foundations to establish a robust framework for understanding the limits of robustness in RL or to apply their findings to real-world problems, such as in autonomous systems, healthcare, finance, and robotics, where reliability and safety are paramount. They will work with a dynamic team of researchers and collaborate with experts in reinforcement learning and trustworthy AI.

Entry Requirements:

Candidates must hold a UK Bachelor degree with a minimum of Upper Second Class honours in Computer Science, Mathematics or a closely related discipline or overseas Bachelor degree deemed equivalent to UK Bachelor (by UK ECCTIS) and achieved a grade equivalent to UK Upper Second Class honours in Computer Science, Mathematics or a closely related discipline;

Project specific enquiries:
This program targets individuals with a strong foundation in machine learning, artificial intelligence, reinforcement learning or related fields who are interested in pushing the boundaries of RL in real-world applications.