Title: Explainable AI for Interacting Autonomous Agents

Type of award	PhD Research Studentship
Department	Engineering
Scholarship Details	Scholarship covers full UK PhD tuition fees and a **tax-free** stipend at the current RCUK rate (£14,777 in 2018/19) plus a top-up from an industrial partner (subject to contracts)
Duration	3 years
Eligibility	Home/UK
Starting Date	October 2018

PhD Topic Background/Description

As AI agents are deployed in complex open-ended environments and regularly interact both with other autonomous systems and with humans, we will need to find ways to judge their trustworthiness. This will require them to explain or justify their decisions and actions to meet fundamental standards of transparency. For instance, recently adopted European Parliament regulations grant individuals the right to explanation of any algorithmic decision "which produces legal effects concerning him or her or similarly significantly affects him or her".

Providing such explanations is straightforward if an autonomous system's decisions are based on a manageable number of rules that can be easily understood by humans. However, optimal decision making for complex environments takes account of multiple interactions to minimise multi-dimensional cost functions, making the results extremely difficult to explain, due to the very high number of alternatives considered. Human understandable rules tend to be low dimensional and consider only small numbers of interactions at a time. This identifies the great importance of the trade-off between the performance of a system and how easy it is to understand the rules governing its behaviour.

This PhD project will explore the trade-off between understandability and optimality for autonomous navigation and collision avoidance. As a point of departure, we will study an existing system of rules such as the maritime collision regulations (COLREGS) or the Rules of the Air. These are both established rulesets governing navigation which are easily understandable and sufficiently flexible to cover a broad range of possible scenarios. They have evolved over time to minimise the risk of collision. However, both sets of rules are designed only for pairwise interactions between agents and hence there are interesting questions concerning their scalability and robustness in multi-interaction scenarios. As the project evolves, it will investigate an appropriate modelling framework (e.g. a language) for expressing the behaviour of autonomous agents and their governing rules. That language must be formal enough to support computer operations, natural enough to be understandable to humans, and sufficiently flexible to allow rules to generalise across different but

related scenarios. The project will study and build on many existing frameworks for autonomous decision-making, such as activity planning, temporal logic, declarative programming languages, action selection, case-based reasoning, or fuzzy control. The key questions to be addressed are then as follows:

- Can we identify safe and effective decision policies for interactions between autonomous systems that are expressible in a high-level/understandable language?
- How close are these optimal rule-based policies to globally optimal policies?
- How does the performance loss vary with the number interactions in the ruleset? For example, are rules for scenarios with three interacting systems much better than pairwise rules?

The project will benefit from collaboration with a team of researchers working on the challenges of autonomous agent deployment in the real world. This will include multidisciplinary work on how such agents would influence and interact with the law, especially important in the context of understanding rules of behaviour.

Candidate Requirements

We are looking for an enthusiastic student with either a First or high 2:1 Honours degree in Mathematics, Computer Science Engineering or a related discipline.

Basic skills and knowledge required:

- Enthusiasm for tackling one of the most pressing challenges in the deployment of autonomous shipping (essential)
- Experience of AI, rule-based systems, fuzzy systems and/or multi-agent systems (beneficial)
- Experience of multi-agent systems, autonomous vehicles (maritime or other domains), and/or human-robot interfaces (beneficial)

Informal enquiries

For informal enquiries please contact Prof Jonathan Lawry J.Lawry@bristol.ac.uk
For general enquiries, please email sceem-pgr@bristol.ac.uk

Application Details

To apply for this studentship submit a PhD application using our online application system [www.bristol.ac.uk/pg-howtoapply]

Please ensure that in the Funding section you tick “I would like to be considered for a funding award from the Engineering Mathematics Department” and specify the title of the scholarship in the “further details” box below with the name of the supervisor Prof Jonathan Lawry.

Apply now