Agent Verification Workshop
- Panel Discussion -

Michael Rovatsos
School of Informatics, The University of Edinburgh

11 September 2015, Liverpool, UK

Previous work on “verification”

- Automated Norm synthesis
  - Deriving behavioural constraints from conflict state specifications in a scalable way
- Argumentation-based conflict resolution
  - Aligning different views about plans based on local viewpoints
- Data mining agent communication
  - Qualitative modelling of trust and reputation, but more generally knowledge-level behaviour prediction

1. *What do you think is the single most exciting or important open problem or unanswered question that the field of Agent Verification should tackle?*

What guarantees we can give about agent behaviour in settings with long-term learning?

- Whether we actually should focus on giving guarantees about adaptation rather than deployment-time
- Safety will be much more important than optimality

2. A lot of research in Agent Verification is being driven by concerns around the certification of autonomous systems. What other areas, do you think, present significant and exciting challenges?

"Human-friendliness" depends on goal-setting, much more work needs to go into that

- Ethical issues depend on values embedded in systems design, interactive/collective process
- Really an issue of autonomous systems governance
- "Limited autonomy" e.g. algorithmic decision making is already out there, shouldn't focus on "visionary" systems

3. Thinking of challenges in terms of risks or dangers, is there anything the field of Agent Verification should be concerned about?

- Need to make sure we focus on autonomy, otherwise “just more verification stuff”
- Avoid focus on existing, partially entrenched, academic communities
- A lot of challenges arise from placing autonomous systems in the real world
- Need to think about semantics and human-machine interface