1. PROJECT TITLE: Synthesis of Trustworthy Behaviour of Artificial Agents (SYBA)

2. AIMS AND BACKGROUND

Nature of the challenge. Systems built on the insights of Artificial Intelligence are increasingly deployed in the world as *agents*, e.g., software agents negotiating on our behalf on the internet, driverless cars, robots exploring new and dangerous environments, bots playing games with humans. There is an obvious need for humans to be able to *trust* the decisions made by artificial agents, the need for meaningful interactions between humans and agents, and the need for transparent agents [2].

This need can only be met if humans are able to model, control and predict the behaviour of agents. This challenge is made all the more complicated since: 1) agents are often deployed with *other* agents leading to *multi-agent systems*, 2) agent behaviour is complex, and extends into the future, leading to *temporal reasoning*, 3) agents are often self-interested with personal goals, leading to *strategic reasoning*, 4) agents may have uncertainty about the state, or even the structure, of other agents and the environment, leading to *epistemic reasoning*.

Focus of this project. Building and analysing artificial agents requires that we do *Temporal-Strategic-Epistemic reasoning on multi-agent systems*. The aim of this project is to develop the mathematical foundations and computational techniques for building and analysing trustworthy artificial agents, by leveraging the insights from recent results developed by the candidate on modeling, control and analysis of single and multi-agent systems.

State of the art. Logic-based techniques are a standard approach to modeling, building and analysing computational systems. Indeed, simply formalising the reasoning tasks unambiguously requires a formal language. Not surprisingly, such reasoning is computationally *undecidable* when it involves epistemic reasoning, a fact known since the late 1970s [45]. The standard approach to ameliorate this is to restrict to classes of multi-agent systems in which agents' private information is hierarchical (typically, one assumes some sort of hierarchy on agent observation or information [45, 46, 33, 12, 11]). Although mathematically elegant, the *applicability of such assumptions is not very high* since in almost all meaningful scenarios, agents' private information is not hierarchical.

Proposed approach. In a remarkable recent discovery [8, 9] the candidate defined and explored a very general class of systems that does not suffer from this long-standing limitation, i.e., the class in which *agent actions are fully observable*. He proved that *Temporal-Strategic-Epistemic reasoning is decidable and not harder than the non-epistemic case*. Many scenarios already fall into this class, e.g., distributed computing and multi-party computation based on broadcast communication [38, 1], multi-player games with public play such as poker [15], e-auctions with public bidding [21].

Moreover, the importance of this recent discovery is that it charts an unanticipated path for applying logic-based methods to *meaningful classes* of artificial agents in a *large variety of fields*, for instance: models of collaborative robot exploration in controlled but dynamic environments [41]; models of cloud manufacturing [24]; models of collusion in e-auctions and auction-based mechanisms [21]; models of social networks that use broadcast communication, and thus also formalisations of *twitter* [20, 39]; models of secure cloud-storage that use data-dispersal [34] and secret-sharing protocols [1]; models of multi-player games in which bidding and play is public, such as poker [15].

The team. The candidate is particularly suited to meet this challenge. His background in mathematical logic and formal methods have allowed him to devise effective conceptual frameworks to address problems in AI [49, 6, 19, 5, 13, 28, 13, 8, 9, 11]. His individual expertise is complemented by his close integration with world experts in logic and automata-based verification and synthesis (M.Y. Vardi), multi-agent systems (M. Wooldridge), and automated planning (H. Geffner and B. Bonet).