Sasha Rubin, Research Statement, December 2017

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.

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 the need for reasoning about *time*, 3) agents are often "self-interested", leading to the need to reason about *strategies*, 4) agents may have uncertainty about the state, or even the structure, of other agents and the environment, leading to the need to reason about *knowledge*.

I approach these needs and questions by developing and applying **formal and logical methods** to modeling and reasoning about multi-agent systems. Moreover, multi-agent systems can be viewed as multi-player games, and thus I use notions from game-theory (e.g., strategies, knowledge, and equilibria) to reason about them. I also pursue more foundational/speculative questions such as "What is synthesis and how should it be formalised?".

1 Current Research — Formal methods for multi-agent systems

I have considered three main aspects of multi-agent systems: 1) incomplete information, 2) imperfect information, and 3) quantitative objectives.

1) Incomplete Information This refers to uncertainty about the environment (i.e., the structure of the game). I have considered two sources of incomplete information for MAS.

First, the *number of agents* may not be known, or may not be bounded a priori. In a series of papers, I have contributed to a generalisation of a cornerstone paper on verification of such systems ("Reasoning about Rings", E.A. Emerson, K.S. Namjoshi, POPL, 1995) from ring topologies to arbitrary topologies [C18],[C19],[C10],[J1]. Other work on this topic studied the relative power of standard communication-primitives assuming an unknown number of agents [C15], as well as the complexity of model-checking timed systems assuming an unknown number of agents [C14]. I also contributed to a book on this topic published by Morgan & Claypool in 2015 [B1]. Recently, I have studied abstraction techniques for verifying consensus algorithms from the distributed computing literature [C1].

Second, the *environment* may not be known, or may be partially-known. For instance, the agents may know they are in a ring, but may not know the size of the ring. I initiated the application of automata theory for the verification of high-level properties of light-weight mobile agents in partially-known environments [C17]. In follow-up work I explored this theme further, including finding ways to model agents on grids — the most common abstraction of 2D and 3D space [C13],[C8],[C16]. I also explored partially-known environments in the context of **automated planning**. Planning in AI can be viewed as the problem of finding strategies in succinct representations of one- or two-player graph-games. In this model vertices represent states, edges represent transitions, and the players represent the agents. I have contributed foundational work to such games. Concretely, I recently extended the classic belief-space construction for games of imperfect-information from finite arenas to infinite-arenas [C11]. Infinite arenas often arise in the study of MAS with incomplete information (for instance, consider a scenario that an agent needs to chop down a tree but does not know how many chops are needed to fell the tree; this incomplete information about the environment is naturally modeled as a single infinite-state arena). I have also used these ideas to elucidate the role of observation-projections in generalised planning problems [C5].

2) Imperfect Information Even if agents have certainty about the structure of the system, they may not know exactly which state the system is in. This is called imperfect information and the associated logics for reasoning about such cases are called *epistemic*. I have studied strategic-epistemic logics in a number of works, namely, with a prompt modality (thus allowing one to express that a property holds "promptly" rather than simply "eventually") [C9], and on systems with public-actions (such as certain card games, including a hand of Poker or a round of Bridge) [C3],[C2]. The importance of these last works is that they give the first decidability (and sometimes optimal complexity) results for strategic reasoning about games of imperfect information in which the agents may have arbitrary observations. In contrast, following classical restrictions on the observations or information of agents, I have also shown how to extend strategy logic by epistemic operators and identified a decidable fragment in which one can express equilibria concepts [C2].

3) Quantitative Objectives I have generalised classic results about certain games with quantitative objectives (i.e., Ehrenfeucht and J. Mycielski. Positional strategies for mean payoff games. International Journal of Game Theory, 8:109–113, 1979) to so-called first-cycle games [J3], thus providing an easy-to-use recipe for deciding if a given agent can use a memoryless strategy to play optimally. I have studied MAS in which agents have a *mix* of qualitative and quantitative objectives [C6], and proved that one can decide if a multi-player game has a Nash Equilibrium in such a setting. Finally, I recently established and studied a logical formalism, called "graded strategy-logic", that is rich enough to *count* equilibria [J2],[C7]. The importance of this result to equilibrium selection is that it gives a computational way to decide if a given game with Boolean objectives has a *unique* Nash equilibrium (and thus supply strong predictions on rational play).

2 Future Research Plans — Building trustworthy agents

As discussed in the introduction, synthesising and analysing trustworthy artificial agents requires *Temporal-Strategic-Epistemic reasoning on Multi-agent Systems*. I plan to develop the mathematical foundations and computational techniques for building and analysing trustworthy artificial agents, by leveraging the insights from my and others' recent results on modeling, control and analysis of single and multi-agent systems. I have three specific **objectives**: 1) discover new classes of systems for which Temporal-Strategic-Epistemic reasoning is decidable and tractable, 2) develop the theory of reasoning about optimal strategies and socially optimal equilibria, and 3) establish scalable algorithms and tools for Temporal-Strategic-Epistemic reasoning.

Proposed approach. In a recent discovery [C2],[C3] we defined and explored a very general class of systems that does not suffer from long-standing limitations [15]. The class is that in which *agent actions are fully observable*. This worked showed 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 [12, 1], multi-player games with public play such as poker [4], e-auctions with public bidding [9]. 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 [14]; models of cloud manufacturing [10]; models of collusion in e-auctions and auction-based mechanisms [9]; models of social networks that use broadcast communication, and thus also formalisations of *twitter* [8, 13]; models of multi-player games in which bidding and play is public, such as poker [4]; models of secure cloud-storage that use data-dispersal [11] and secret-sharing protocols [1].

In order to meet objective 1) I propose to generalise systems in which all actions are fully observable, as well as explore orthogonal systems to achieve decidability and tractability; in order to meet objective 2) I propose to enrich the models and specification languages with costs/rewards and analyse these with new measures of strategy quality [2, 5, 6]; in order to meet objective 3) I propose to translate Temporal-Strategic-Epistemic reasoning to Automated Planning extending and refining existing translations [3, 16, 7].

3 Past Research — Algorithmic Model Theory

My prior work contributed to a research program called "Algorithmic Model Theory" whose aim is to develop and extend the successes of Finite Model Theory to infinite structures that can be reasoned about algorithmically. Specifically, my PhD work pioneered the development of "automatic structures": this is a generalisation of the regular languages from sets to mathematical objects with structure, such as graphs, arithmetics, algebras, etc. The fundamental property of automatic structures is that one can automatically answer logic-based queries about them (precisely, their first-order theory is decidable). I gave techniques for proving that structures are or are not automatic (similar to, but more complicated than, pumping lemmas for regular languages), I studied the computational complexity of deciding when two automatic structures are the same (isomorphic), and I found extensions of the fundamental property, thus enriching the query language [BC1],[C30],[C27],[J6],[C29],[C28],[J7],[J5]. Finally, I have also worked on extensions of automatic structures to include oracle computation [C21],[C22].

References

[1] I. Abraham, D. Dolev, R. Gonen, and J. Halpern. Distributed computing meets game theory: Robust mechanisms for rational secret sharing and multiparty computation. In *PODC06*, pages 53–62, New York, NY, USA, 2006. ACM.

[2] S. Almagor, U. Boker, and O. Kupferman. Formally reasoning about quality. *J. ACM*, 63(3):24:1–24:56, June 2016.

[3] J. A. Baier and S. A. McIlraith. Planning with first-order temporally extended goals using heuristic search. In *AAAI'06*, pages 788–795. AAAI Press, 2006.

[4] M. Bowling, N. Burch, M. Johanson, and O. Tammelin. Heads-up limit hold'em poker is solved. *Science*, 347(6218):145–149, 2015.

[5] R. I. Brafman, G. De Giacomo, and F. Patrizi. Specifying non-markovian rewards in mdps using LDL on finite traces (preliminary version). *CoRR*, abs/1706.08100, 2017.

[6] T. Brihaye, G. Geeraerts, A. Haddad, and B. Monmege. Pseudopolynomial iterative algorithm to solve total-payoff games and min-cost reachability games. *Acta Inf.*, 54(1):85–125, 2017.

[7] A. Camacho, E. Triantafillou, C. Muise, J. Baier, and S. McIlraith. Non-deterministic planning with temporally extended goals: LTL over finite and infinite traces.

[8] R. De Nicola, A. Maggi, M. Petrocchi, A. Spognardi, and F. Tiezzi. Twitlang(er): Interactions modeling language (and interpreter) for twitter. In R. Calinescu and B. Rumpe, editors, *SEFM'15*.

[9] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* Cambridge, 2010.

[10] P. Felli, L. de Silva, B. Logan, and S. M. Ratchev. Process plan controllers for non-deterministic manufacturing systems. In *IJCAI'17*.

[11] M. Li, C. Qin, J. Li, and P. P. C. Lee. Cdstore: Toward reliable, secure, and cost-efficient cloud storage via convergent dispersal. *IEEE Internet Computing*, 20(3):45–53, 2016.

[12] N. A. Lynch. *Distributed Algorithms.* Morgan Kaufmann, 1996.

[13] A. Maggi, M. Petrocchi, A. Spognardi, and F. Tiezzi. A language-based approach to modelling and analysis of twitter interactions. *J. Log. Algebr. Meth. Program.*, 87:67–91, 2017.

[14] C. Newcombe, T. Rath, F. Zhang, B. Munteanu, M. Brooker, and M. Deardeuff. How Amazon web services uses formal methods. *Communications of the ACM*, 58(4):66–73, 2015.

[15] A. Pnueli and R. Rosner. Distributed reactive systems are hard to synthesize. In *FOCS'90*, pages 746–757, 1990.

[16] J. Torres and J. Baier. Polynomial-time reformulations of LTL temporally extended goals into final-state goals. In *IJCAI*, pages 1696–1703, 2015.