Research plan 2018-2021

Sasha Rubin

1 Context

My primary interest is in **Formal Methods for Artificial Agents**.

Formal methods (FM) is an umbrella-term that describes principles and techniques for reasoning about systems with some digital component, such as software, hardware, cyber-physical systems, etc.

> FM is built on three pillars: modeling, verification, and synthesis.

All three pillars are founded on techniques from mathematical logic. Indeed, logic is used to formally represent aspects of the system so that both (expert) humans and computers can manipulate and reason about them.

Artificial Intelligence (AI) studies the principles behind thinking and reasoning, and it does so in a mathematical and algorithmic way. *Agents* are entities that can interact with each other and their environment using sensors and actuators [17]. Agents built on the insights of AI are called *artificial agents (AA)*. As artificial agents are increasingly deployed in the world (e.g., software agents on the Internet, driverless cars, software that play and compete with humans at games, robots exploring new and dangerous environments, etc.) a clear challenge has materalised: there is a need for humans to be able to *trust* the decisions made by AA, the need for *meaningful interactions* between humans and AA, and the need for *transparent* AA [1]. One path to meeting these needs is to create *explainable AI (XAI)*, i.e., to enable humans to understand, trust and manage systems built using AI [9].

This is a grand challenge that involves many facets of computer science: psychology (to understand what is a good explanation), knowledge representation and logic (to formalise this in a way accessible to humans and machines), algorithms (to produce good explanations), NLP (to interface with human users), software engineering (to produce reliable and efficient programs), etc.

I take the following as an hypothesis:

> The challenge of producing transparent and explainable artificial agents (AA) cannot be met without having some formal guarantees on the behaviour of AA.

It is in this context that formal methods will play a role: “Modeling” problems at the correct levels of abstraction to be able to reason about them, “Verification” for deducing that AA do indeed have specified properties, and “Synthesis” for automatically producing correct-by-design AA.

2 What is an environment?

An *environment* provides the surrounding conditions for agents to exist and operate [12]. An environment is tri-modal: it consists of a **physical-environment** that supports the agents physically (e.g., physical roads, network links); a **communication-environment** that supports agent communication (e.g., rules and protocols, pheromones); and a **social-environment** that reflects organisational structure in terms of roles, groups, etc.

Here are two scenarios that describe artificial agents in an environment.

Example 1 (Multiplayer trick-based card-games)**.** *Let us consider one way to model Bridge. The physical-environment entails that the environment deals the cards so that no human agent can see the cards held by the other agents. The social-environment consists of players' teams, and a team's objective is to win more tricks than the opposition. The communication-environment ensures that agents are required to talk publicly in the bidding stage (and thus learn more information about what cards other agents may have), while private talking is prohibited.*

Example 2 (Mobile agents)**.** *Mobile agents can be modeled as multiple agents in a discrete world. The physical environment consists of the common space such as a grid or a graph in which the agents move. The communication environment may enforce that agents only have line of sight, or they may have RDF-sensors, etc. The social environment may include teams (e.g., for finding a missing person), or adversarial individuals (e.g., get to the finish line first).*

In the next 4 years, I plan to focus on one central piece of the problem of creating AA that humans can understand, trust and manage:

Motivating Problem. *When an AA interacts with the world, which may include humans and other AAs, it is situated in an* ***environment*** *[17]. How should the environment be abstracted, modeled and exploited in the context of formal methods?*

I know show that formal-methods models **environments in an oversimplistic way**. Classically, the FM literature considers the environment to be an amalgam of everything not under the control of the designer, see [8, 13, 18]. This view is overly simplistic as it ignores the multi-modal nature of the environment. Interestingly, the multi-agent systems (MAS) community has established that it is both essential and natural to treat the **environment as a first-class citizen** [19, 12]. However, **this was mainly explored in the context of MAS architectures, with no formal methods or rigorous mathematical results.** Indeed, FM continues to treat the environment as an amalgam of all its different aspects that is either modeled as another agent or described as a logical theory/formula [13, 15]. Unsurprisingly, this has a number of disadvantages:

1. it is a non-trivial modeling task to specify and model this way all possible "behaviors" or relevant aspects of the environment [10];
2. agents are used to provide functionalities and services that are not appropriate for them [19];
3. all aspects of the environment are entangled with each other and in many cases also with in the state-space of the whole system, making reasoning about it difficult [19].

Note that while treating the environment as another agent may appear as giving it "first class" status, it still ignores the fact that the environment has very different characteristics than an agent: it need not have any goals, and it need not behave strategically; modeling incomplete information of the agents regarding the environment is very cumbersome; and finally, since simple aspects of the environment are modeled with the powerful concept of an agent (e.g., the ability to communicate privately with other agents), one can easily end up with an undecidable or intractable model-checking or synthesis problem, even for finite-state systems and strategies [11, 14].

In recent years, the above concern that an environment need not be purely collaborative or purely adversarial has lead to the use of multiplayer games in which agents may have overlapping objectives, i.e., non-zero sum games. However, even in this setting, the rest of the deficiencies in modeling the environment remain.

3 Aim and Impact

I will develop new formal methods in which environments are separated into their three components (physical, communication, social).

In particular, I plan to define parameters of each of the components and prove theorems that show how the computational complexity of verification and synthesis tasks changes as one varies the parameters. Isolating such parameters will allow for a fine-grained view of the borders of tractability and decidability for verification and synthesis tasks in formal-methods. An immediate payoff of such a theory of the environment is that one will be able to exploit the theorems, algorithms and insights that result in order to verify and synthesise agents that were previously out of reach.

The separation will provide the leverage to develop **meaningful models** and **tractable problems**.

4 Approach and connection to previous work

I propose a two-stage approach. First, identify relevant parameters of each of the components. Second, devise provably optimal algorithms for verification and synthesis tasks for the parameterised systems. In particular, by isolating the environment components, one can achieve decidability of problems that are undecidable when considering the environment as a single amalgam. The following preliminary works shows the feasability of this approach.

1. I have considered the the physical component in my recent Marie-Curie COFUND project on verification of mobile agents in discrete but partially known environments [16, 2, 3]. This work models the physical-environment as a graph and parameterises these graphs by a natural width-parameter (i.e., clique-width).
2. I have considered the communication component in the verification of MAS by isolating a class of systems in which agents have imperfect information and communicate by broadcast [4, 6, 5].
3. I have considered the social component in recent studies of epistemic extensions of strategic logics [7, 5]. These works show that one can achieve decidability by carefully restricting the information that agents have about each other.

The purpose of this project is to build upon these preliminary works and establish the environment as a first-class citizen in the formal-methods literature. Moreover, these works only deal with decidability; a major aspect of this project will be to find meaningful parameters and values (i.e., that correspond to natural problems) that also yield tractable, or at least low computational complexity, verification and synthesis tasks.

References

[1] ACM U.S. Public Policy Council and ACM Europe Policy Committee. Statement on algorithmic transparency and accountability. ACM, 2017.

[2] B. Aminof, A. Murano, S. Rubin, and F. Zuleger. Verification of asynchronous mobile-robots in partially-known environments. In *PRIMA 2015: Principles and Practice of Multi-Agent Systems - 18th International Conference, Italy, 2015, Proceedings*, pages 185–200, 2015.

[3] B. Aminof, A. Murano, S. Rubin, and F. Zuleger. Automatic verification of multi-agent systems in parameterised grid-environments. In *Intl. Conf. on Autonomous Agents and Multiagent Systems (AAMAS)*, pages 1190–1199, 2016.

[4] B. Aminof, S. Rubin, and F. Zuleger. On the expressive power of communication primitives in parameterised systems. In *Logic for Programming, Artificial Intelligence, and Reasoning - 20th International Conference, LPAR-20 2015, Suva, Fiji, November 24-28, 2015, Proceedings*, pages 313–328, 2015.

[5] Francesco Belardinelli, Alessio Lomuscio, Aniello Murano, and Sasha Rubin. Verification of broadcasting multi-agent systems against an epistemic strategy logic. In *International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 2017.

[6] Francesco Belardinelli, Alessio Lomuscio, Aniello Murano, and Sasha Rubin. Verification of multi-agent systems with imperfect information and public actions. In *Proceedings of the 2017 International Conference on Autonomous Agents & Multiagent Systems (AAMAS 2017)*, 2017.

[7] Raphael Berthon, Bastien Maubert, Aniello Murano, Sasha Rubin, and Moshe Vardi. Hierarchical strategic reasoning. In *IEEE Symposium on Logic in Computer Science (LICS 2017)*, 2017.

[8] R. Fagin, J.Y. Halpern, Y. Moses, and M.Y. Vardi. *Reasoning about Knowledge.* MIT Press, 1995.

[9] David Gunning. Explainable artificial intelligence (xai). Broad Agency Announcement. DARPA-BAA-16-53.

[10] Hai Lin. Mission accomplished: An introduction to formal methods in mobile robot motion planning and control. *Unmanned Systems*, 2, 2014.

[11] Fabio Mogavero, A. Murano, Giuseppe Perelli, and Moshe Y. Vardi. Reasoning about strategies: On the model-checking problem. *ACM Trans. Comput. Log.*, 15(4):34:1–34:47, 2014.

[12] James Odell, H. Van Dyke Parunak, Mitchell Fleischer, and Sven Brueckner. Modeling agents and their environment. In Fausto Giunchiglia, James Odell, and Gerhard Weiß, editors, *Agent-Oriented Software Engineering III, Third International Workshop, AOSE 2002, Bologna, Italy, July 15, 2002, Revised Papers and Invited Contributions*, volume 2585 of *Lecture Notes in Computer Science*, pages 16–31. Springer, 2002.

[13] A. Pnueli and R. Rosner. On the synthesis of a reactive module. In *POPL*, pages 179–190, 1989.

[14] Amir Pnueli and Roni Rosner. Distributed reactive systems are hard to synthesize. In *31st Annual Symposium on Foundations of Computer Science, St. Louis, Missouri, USA, October 22-24, 1990, Volume II*, pages 746–757. IEEE Computer Society, 1990.

[15] Raymond Reiter. *Knowledge in action: logical foundations for specifying and implementing dynamical systems.* MIT press, 2001.

[16] S. Rubin. Parameterised verification of autonomous mobile-agents in static but unknown environments. In *AAMAS*, pages 199–208, 2015.

[17] Stuart J. Russell and Peter Norvig. *Artificial Intelligence - A Modern Approach (3. internat. ed.).* Pearson Education, 2010.

[18] H. Veith R. Bloem T. Henzinger, E. Clarke, editor. *Handbook of Model Checking.* 2017.

[19] Danny Weyns, Andrea Omicini, and James Odell. Environment as a first class abstraction in multiagent systems. *Autonomous Agents and Multi-Agent Systems*, 14(1):5–30, 2007.