Formal Methods for Automated Synthesis of Trustworthy Behaviour of Artificial Agents (TRUST)

Even if you are not planning to do implementation yourself, talking about a collaboration with others who are doing that might help.

structure the first page so that it reads as an executive summary of the proposal that conveys to a *non-expert* reader an overview of what it is about,

Part A

Summary (max 750 chars) AI systems are increasingly deployed in the world as *agents*, e.g., software negotiating on our behalf on the internet, driverless cars, robots exploring dangerous environments, etc. There is a recently articulated need for humans to be able to *trust* the decisions made by such artificial agents [2]. The goal of this project is to develop mathematical foundations and computational techniques for building trustworthy artificial agents, by leveraging the insights from recent results, developed by the proposer, on synthesis and strategic reasoning for single and multi-agent systems. The projected benefit will be techniques, useable by computer scientists and engineers, for building trustworthy agents in realms such as high-level robot control including lightweight swarms, concurrent manufacturing in industry 4.0, trustworthy social-media and -news bots, safe and secure could storage facilities.

Benefit and Impact (max 750 chars) The anticipated benefit of this project to science is that it will advance the state-of-the-art of the verification and synthesis of artificial agents. The potential impact to UNSW's Strategic Theme "Future Intelligence" will be tighter integration with world-renowned experts in AI and Autonomous Systems, including attracting short- and long-term leaders in Automated Planning and Knowledge Representation. Safer and securer interactions with artificial agents are in the interest of society.

Part C (max 10 pages)

PROJECT TITLE Formal Methods for Automated Synthesis of Trustworthy Behaviour of Artificial Agents (TRUST)

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

In other words, humans must be able to model, control and predict the *behaviour* of agents. This challenge is made all the more complicated since:

- agents are often deployed with other agents leading to *multi-agent systems*,
- agent behaviour is complex, and extends into the future, leading to *temporal reasoning*,
- agents are required to behave strategically, leading to *strategic reasoning*,
- agents may have uncertainty about the state, or even the structure, of other agents and the environment, leading to *epistemic reasoning*.

The **aim** of this project is to develop mathematical foundations and computational techniques for building trustworthy artificial agents, by leveraging the insights from recent results (developed by the candidate) on synthesis and strategic reasoning for single and multi-agent systems.

Specifically:

1. We need to discover meaningful *new classes* of agents for which temporal-strategic-epistemic reasoning is decidable and tractable.

It is known that synthesis for systems composed of multiple agents having imperfect information is **undecidable** (this has been discovered in the multiple contexts, i.e., decentralised POMDPs [5], multiplayer non-cooperative games of imperfect information [24], distributed synthesis [25]). Since the 1990s researchers have tried to find meaningful decidable fragments. The standard approach is to assume some sort of hierarchy on the information or observation sets, e.g., [6]. Although mathematically elegant, the applicability of such assumptions is not very high. However, we recently defined and explored a class of games in which all agent moves are public, and proved that one can do analysis, i.e., the model-checking problem for such games against temporal-strategic-epistemic logics is **decidable and not harder than the case of perfect information** [C2],[C1] [1] (previously it was only known that, in a similar setting, one can do multi-agent epistemic planning [18] and synthesis [31]). The importance of this result is that it lays the algorithmic and theoretical foundations for analysing temporal-strategic-epistemic properties of *meaningful* classes of agents such as:

- various models of distributed computing using broadcast communication, and thus also formalisations of *twitter* [10, 21].
- various models of secure cloud-storage that use data-dispersal [19] and secret-sharing protocols [1].
- various multi-player games such as poker, stratego and bridge [**?**].

2. We need to reduce temporal-strategic-epistemic reasoning about agents to *scalable tools* such as classical and fully-observable non-deterministic planners developed in the Planning in AI community.

[1] References in the form [C##], [J##] and [B##] refer to conferences, journals and books by the candidate, and can be found in the "References" section of the accompanying CV.

Planning is a branch of AI that addresses the problem of generating a course of action to achieve a desired goal, given a description of the domain of interest and its initial state. **Planning is a form of synthesis** that is central to the development of agents. Besides theoretical insights, Planning provides practical tools based on heuristic search and symbolic methods [13]. The most successful of this technology is for "classical planning", i.e., single agent, deterministic environment, with perfect information, and simple reachability goals, and "fully observable non-deterministic planning" (which amounts to the case of one agent in an adversarial environment). Previous work has reduced planning with temporal goals to classical and fully-observable nondeterministic planning [4, 29, 8]. This lays the foundation for refining and extending the translations to handle temporal-strategic-epistemic reasoning.

3. We need to define, analyse, and tackle the problem of synthesising *optimal strategies* in systems of agents with *quantitative objectives*.

Previous work in planning focused on optimal strategies for MDPs and POMDPs, as well as optimal plans [13, 23, 28]. However, reasoning about multiple agents with possibly different but overlapping objectives requires richer solution concepts from **Game Theory**, e.g., Nash Equilibrium and Pareto optimality. The candidate recently introduced expressive temporal-strategic-epistemic logics that can be used to reason about equilibria in cases agents have Boolean objectives [C1]. This, together with recent insights from quantitative verification [3], lays the foundation for designing useful logics that can reason about systems in which agents have quantitative objectives.

FUTURE FELLOWSHIP CANDIDATE I have deep and extensive knowledge in logics for temporal, strategic and epistemic reasoning, automata-theory and synthesis. I have contributed foundational work on synthesis and graph-games [J5],[J7] as well as on the connections between synthesis and general forms of planning [C10],[W1],[C4].

A first step towards synthesis is usually verification, and I have contributed deep work on verification of multi-agent systems [C17],[C18],[C13],[C9],[J1], including a book on the topic [B1],[J6], and with a focus on verification of parameterised systems [C16],[C12],[C7].

I already have close connections with international experts with expertise and interest in the topic of this project, i.e., including Giuseppe De Giacomo (knowledge representation, artificial intelligence, verification, synthesis), Hector Geffner and Blai Bonet (Planning), Moshe Vardi and Aniello Murano (logics for strategic reasoning, automata-theory for synthesis and verification).

PROSPOSED PROJECT AND QUALITY INNOVATION The challenge of building trustworthy agents cannot be met without having some formal guarantees on their behaviour. The holy-grail is to automatically synthesise agent behaviour, or part of their behaviour, from specifications. This project will advance the state of the art of the mathematical foundations and computational techniques for building trustworthy agents from temporal-strategic-epistemic specifications. talk about specific outcomes

Objectives and methodology. The objectives of the project are to generate new mathematics, algorithms, and pragmatic techniques for describing, reasoning-about and building trustworthy agents. This will be done using methods and insights from Logic and Formal Methods (and program synthesis in particular), and Game Theory (and its development in multi-agent systems).

For instance:

1. in order to get new and richer decidable classes, I propose to *generalise* systems in which all moves are public (which we recently explored [C2],[C1]), e.g., by incorporating stochastic initial states, by allowing a bounded number of private moves.
2. in order to achieve tractable classes of agents, I propose to restrict systems in which all moves are public to homogenous initial conditions [20] and/or bounded-epistemic states.
3. in order to reason about optimal strategies, I propose to enrich the models with costs/rewards and analyse these with new measures of strategy-quality.

FEASABILITY AND STRATEGIC ALIGNMENT ï· Describe the extent to which the Future Fellowship Candidate aligns with and/or complements the core or developing research strengths and staffing profile of the Administering Organisation. ï· Demonstrate that the necessary facilities are available to conduct the proposed research. ï· Outline what resources will be provided by the Administering Organisation to support the Future Fellowship Candidate during her/his Future Fellowship. ï· At the end of the Future Fellowship, explain what capacity exists at the Administering Organisation to transition the Candidate to a continuing position.

BENEFIT AND COLLABORATION This project relies on the hypothesis that, in order to synthesise and verify trustworthy agents, one needs to buid a model of the agent. The model-based approach to controller design underlies automated planning in AI [13], reactive synthesis in programming verification [7], generalised game playing [14], Bayesian networks and decision graphs [17], etc. This is in contrast to the model-free approach as epitomised by the recent commercial application of neural networks to vision problems. The key issue in AI is not whether we should use the model-based or the model-free approach. Rather, the challenges are to gain *understanding* of the limitations of each technique.

Thus, the potential impact of this project is that it will *advance our scientific understanding of the reach and limitations of the model-based approach* for designing and reasoning-about the behaviour of artificial agents.

COMMUNICATION OF RESULTS ï· Outline plans for communicating the research results to other researchers and the broader community, including but not limited to scholarly and public communication and dissemination

MANAGEMENT OF DATA ï· Outline plans for the management of data produced as a result of the proposed research, including but not limited to storage, access and re-use arrangements. ï· It is not sufficient to state that the organisation has a data management policy. Researchers are encouraged to highlight specific plans for the management of their research data.

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

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

[4] 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.

[5] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein. The complexity of decentralized control of markov decision processes. *Math. Oper. Res.*, 27(4):819–840, 2002.

[6] D. Berwanger, A. B. Mathew, and M. van den Bogaard. Hierarchical information patterns and distributed strategy synthesis. In B. Finkbeiner, G. Pu, and L. Zhang, editors, *ATVA'15*, volume 9364 of *Lecture Notes in Computer Science*, pages 378–393. Springer, 2015.

[7] R. Bloem, B. Jobstmann, N. Piterman, A. Pnueli, and Y. Sa'ar. Synthesis of reactive(1) designs. *J. Comput. Syst. Sci.*, 78(3):911–938, 2012.

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

[9] F. Cassez, R. van der Meyden, and C. Zhang. The complexity of synchronous notions of information flow security. *Theor. Comput. Sci.*, 631:16–42, 2016.

[10] 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*.

[11] S. Eggert, R. van der Meyden, H. Schnoor, and T. Wilke. The complexity of intransitive noninterference. In *S&P'11*, pages 196–211. IEEE Computer Society, 2011.

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

[13] H. Geffner and B. Bonet. *A Concise Introduction to Models and Methods for Automated Planning*. Morgan & Claypool Publishers, 2013.

[14] M. Genesereth and M. Thielscher. General game playing. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 8(2):1–229, 2014.

[15] X. Huang and R. van der Meyden. Symbolic model checking epistemic strategy logic. In C. E. Brodley and P. Stone, editors, *AAAI'14*, pages 1426–1432. AAAI Press, 2014.

[16] X. Huang and R. van der Meyden. The complexity of approximations for epistemic synthesis (extended abstract). In *SYNT'15*, pages 120–137, 2015.

[17] F. V. Jensen. *Bayesian Networks and Decision Graphs*. Information Science and Statistics. Springer, 2002.

[18] F. Kominis and H. Geffner. Beliefs in multiagent planning: From one agent to many. In R. I. Brafman, C. Domshlak, P. Haslum, and S. Zilberstein, editors, *ICAPS'15*, pages 147–155. AAAI Press, 2015.

[19] 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.

[20] A. Lomuscio, R. van der Meyden, and M. Ryan. Knowledge in multiagent systems: initial configurations and broadcast. *ACM Trans. Comput. Log.*, 1(2):247–284, 2000.

[21] 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.

[22] A. K. McIver and C. C. Morgan. Compositional refinement in agent-based security protocols. *Formal Aspects of Computing*, 23(6):711–737, Nov 2011.

[23] G. D. Penna, B. Intrigila, D. Magazzeni, and F. Mercorio. Synthesis of cost-optimal strong plans in non-deterministic domains. *International Journal on Artificial Intelligence Tools*, 24(06):1550025, 2015.

[24] G. Peterson, J. Reif, and S. Azhar. Lower bounds for multiplayer noncooperative games of incomplete information. *Computers & Mathematics with Applications*, 41(7-8):957–992, 2001.

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

[26] D. Rajaratnam, B. Hengst, M. Pagnucco, C. Sammut, and M. Thielscher. *Composability in Cognitive Hierarchies*, pages 42–55. Springer, Cham, 2016.

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

[28] Á. Torralba, V. Alcázar, P. Kissmann, and S. Edelkamp. Efficient symbolic search for cost-optimal planning. *Artif. Intell.*, 242:52–79, 2017.

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

[30] R. van der Meyden and M. Y. Vardi. Synthesis from knowledge-based specifications (extended abstract).

In *CONCUR'98*, pages 34–49, 1998.

[31] R. van der Meyden and T. Wilke. Synthesis of distributed systems from knowledge-based specifications. In *05*, LNCS 3653, pages 562–576, 2005.

Integration at UNSW There are clear connections between the topic of this project (behaviour synthesis for agents) and the work being done at UNSW.

In particular:

1. The theory of Reasoning about Knowledge, as applied to distributed computing [12], can be used to formally specify, verify and synthesise artificial-agents that act with incomplete and imperfect information. **Van der Meyden** studies synthesis of epistemic protocol specifications [30, 16] and symbolic implementations of model-checkers for epistemic strategic logics [15].
2. General game playing, see **Thieschler [14]**, is a framework in which programs learn to play games given just a description of the rules of the game. Although related to synthesis, it emphasises an online approach in which the solver is given bounded time to suggest its next move (in this sense it is related to online planning, e.g., [13]). This is in contrast to the classic synthesis approaches in Formal Methods which are offline and generate a policy before execution.
3. Architectures used in robotics to capture mental states correspond to a first-person view of agents [27]. Recent work at UNSW aims to formalise an architecture so that formal-methods can be applied, see **Rajaratnam, Hengst, Pagnucco, Sammut, Thielscher [26]**.
4. The aim in information flow security is to design observable program behaviour that does not reveal, to an adversary, secret information about its' state. Formal methods for the design and verification of such systems is studied by **Morgan [22]** and **van der Meyden [11, 9]**. Definitions of information flow security serve as important specification of multi-agent system behaviour, e.g., in secure multi-party communication.
5. Walsh? Aziz?