Verifiable Autonomy using Rational Agents

Louise A. Dennis[1] Maryam Kamali[1] Michael Fisher[1]

Department of Computer Science, University of Liverpool, L.A.Dennis@liverpool.ac.uk

Abstract: We are interested in the verification of autonomous systems that use a *rational agent* to make decisions. We discuss the use of model-checking to provide guarantees about the behaviour of such systems.

1 Introduction

Hybrid autonomous systems combine low-level control algorithms with high level reasoning techniques from artificial intelligence. These systems control machines such as cars, drones and robots that move and act on the physical world. Governments, industry and the public anticipate rapid advances in these areas over the coming decades, particularly in the technologies for driverless cars, assistive robots and unmanned aircraft and it is hoped that these technologies will enrich the lives of many people, improve safety and help alleviate the problems of an ageing population. However the public is also anxious about such systems and, in particular, the safety of such systems and their potential capacity to make *stupid* decisions. For this reason, the verification and validation of autonomous systems is an area of active research.

In control engineering, autonomous systems are typically described in terms of their *sensors* and their *actuators*. Sensors provide the system with information about the state of the external world such as temperature, speed, the distance to any obstacle and so on. Actuators control the system's motors and, ultimately, its behaviour in the physical world. Control engineering has developed many algorithms which allow information from sensors to be used to determine the behaviour of the actuators, often when controlling difficult physical behaviours such as allowing a drone to hover in position, or a robot to ride a bicycle. Symbolic Artificial intelligence techniques are used when the system needs to expand beyond single activities, to situations that involve making choices or combining sequences of behaviours.

One technology for achieving this is rational agent programming in which a decision problem is framed in terms of the *beliefs* and *desires* of the system [9]. A rational agent selects programmer supplied *plans* for execution based on these beliefs and desires. In an autonomous system, the beliefs are derived from the information supplied by its sensors, the desires are goals supplied by its users or programmers and the plans are described in terms of sequences of *actions* which generally relate to algorithms in the underlying control system – for instance following a flight path.

Fig. 1 provides a high level view of such a system.

2 Verification of a Single Decision Making Component

Given a system with an architecture similar to that in Fig. 1, we can seek to verify the operation rational-agent based decision making component in isolation [7, 4]. This is motivated partly by the difficulty involved in precisely reasoning about continuous behaviour but also by the observation that the decision-making component is the novel part of many of these systems.

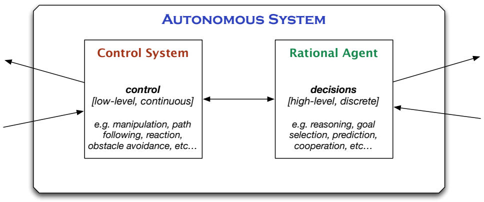
Autonomous System
Control System
control
[low-level, continuous]
e.g. manipulation, path following, reaction, obstacle avoidance, etc…
Rational Agent
decisions
[high-level, discrete]
e.g. reasoning, goal selection, prediction, cooperation, etc…

Fig. 1: A Rational Agent based Hybrid Autonomous System

We are primarily interested in the use of model-checking to verify that all executions of this component obey any safety parameters. We have explored the use of *program model-checking* using the AJPF system [5]. This lets us verify the actual code used to program the system which is a particularly attractive option when dealing with certification issues.

In order to restrict our verification to just the decision-making component we consider all possible sets of beliefs/perception that the agent may hold at each point in time. We show that these always lead to the selection of appropriate action by the agent. What we can not verify is that the beliefs were a correct representation of the real world, nor that the selected action has the desired effect. In effect we verify that correct decisions are made given the information available, but we do not verify the results of those decisions nor the veracity of the information.

3 Verifying Interacting Decision Making Components

While the decision-making components often contains the main novelty of an autonomous system it is important not to underestimate the effect this may have on overall system behaviour both in terms of a single autonomous system and in situations where multiple autonomous systems interact.

To investigate this we have considered a vehicle platooning scenario in which several autonomous cars attempt to form and maintain a platoon behind a car controlled by a human driver. As well as formally verifying individual agent/vehicle decisions, we also represented this system in the UPPAAL model-checker [2] which, in particular, allows the user to explore the real-time properties of a system. To do this we abstract away from the code that programs the rational agents and represent their behaviour in a simple protocol-like form which assumes the correct execution of the individual agent programs.

We are then able to investigate whether the whole convoy is able to meet various timing requirements. For instance, given assumptions about the time taken to change lane, and the time for requests to be made and acknowledgments to be received we can verify that the time between a vehicle requesting entry to a platoon and it assuming its correct place within the platoon falls within acceptable time bounds [8].

4 Verification of Ethical Governors

Model-checking does not scale well as systems and choices increase. This is of concern in applications involving planning and scheduling (and, potentially, learning). Here we may prefer to have a smaller tractable rational agent based component concerned only with reasoning about parts of the execution which have an ethical dimension.

For this we look at the idea of *ethical governors* [1]. We view an ethical governor as a component that can act to filter, prioritise or modify the plans or actions proposed by an underlying autonomous system. It does this in order to conform to ethical considerations. This type of architecture is shown in Fig. 2.

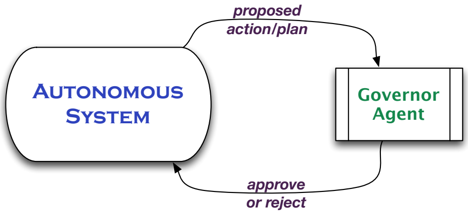

Fig. 2: A Governor Agent monitoring an Autonomous System

In this work the underlying system generates sets of plans or actions and passes these to an Ethical Governor. The governor evaluates the ethical outcomes of these plans or actions and returns either the most ethical or some set of ethically acceptable choices. We model the ethical governor as a rational agent and this allows us to use model-checking to verify the logic used by the ethical governor in order to ensure that, for instance, it only chooses an option in which a human is hurt if all other options had ethically worse outcomes [3, 6].

5 Conclusion

This abstract has surveyed work on the verification of autonomous systems. It has focused on the verification of systems which use a rational agent to make key decisions either in general, or specifically as part of ethical reasoning. We have focused primarily on the verification of these rational agents considered separately from the wider autonomous system but have also discussed preliminary work on how properties of overall system behaviour can be verified.

Acknowledgments

The work in this paper was funded by EPSRC grant Verifiable Autonomy, EP/L024845/1. `http://wordpress.csc.liv.ac.uk/va/`

Data

The case studies described in this abstract can be found at `http://mcapl.sourceforge.net` and `github.com/VerifiableAutonomy`.

References

[1] R.C. Arkin, P. Ulam, and A.R. Wagner. Moral decision making in autonomous systems: Enforcement, moral emotions, dignity, trust, and deception. *Proceedings of the IEEE*, 100(3):571 –589, 2012.

[2] Johan Bengtsson, Kim G. Larsen, Fredrik Larsson, Paul Pettersson, and Wang Yi. UPPAAL — a Tool Suite for Automatic Verification of Real–Time Systems. In *Proc. of Workshop on Verification and Control of Hybrid Systems III*, number 1066 in Lecture Notes in Computer Science, pages 232–243. Springer–Verlag, October 1995.

[3] Louise Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems*, pages –, 2015.

[4] Louise A. Dennis, Michael Fisher, Nicholas K. Lincoln, Alexei Lisitsa, and Sandor M. Veres. Practical verification of decision-making in agent-based autonomous systems. *Automated Software Engineering*, pages 1–55, 2014.

[5] Louise A. Dennis, Michael Fisher, Matthew Webster, and Rafael H. Bordini. Model Checking Agent Programming Languages. *Automated Software Engineering*, 19(1):5–63, 2012.

[6] Louise A. Dennis, Michael Fisher, and Alan F. T. Winfield. Towards verifiably ethical robot behaviour. In *AAAI Workshop on AI and Ethics (1st International Conference on AI and Ethics)*, Austin, TX, January 2015.

[7] Michael Fisher, Louise A. Dennis, and Matthew Webster. Verifying Autonomous Systems. *ACM Communications*, 56(9):84–93, 2013.

[8] M. Kamali, L. A. Dennis, O. McAree, M. Fisher, and S. M. Veres. Formal Verification of Autonomous Vehicle Platooning. *ArXiv e-prints*, February 2016. Under Review.

[9] A. S. Rao and M. P. Georgeff. An Abstract Architecture for Rational Agents. In *Proc. International Conference on Knowledge Representation and Reasoning (KR&R)*, pages 439–449. Morgan Kaufmann, 1992.