Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Formal verification of ethical choices in autonomous systems

Louise Dennis [a,*], Michael Fisher [a], Marija Slavkovik [b], Matt Webster [a]

[a] Department of Computer Science, University of Liverpool, United Kingdom
[b] Department of Information Science and Media Studies, University of Bergen, Norway

HIGHLIGHTS

- An autonomous system should act ethically, but what if it has no all-ethical choice?
- We model how to rank states violating multiple instances of ethical principles.
- We enable an autonomous system to use this ethic rank to rank its available plans.
- We guarantee that when a plan is chosen, it is the most ethical plan available.

ARTICLE INFO

Article history:
Available online 11 December 2015

Keywords:
Autonomous systems
Ethics
BDI programs
Formal verification

ABSTRACT

Autonomous systems such as unmanned vehicles are beginning to operate within society. All participants in society are required to follow specific regulations and laws. An autonomous system cannot be an exception. Inevitably an autonomous system will find itself in a situation in which it needs to not only choose to obey a rule or not, but also make a complex ethical decision. However, there exists no obvious way to implement the human understanding of ethical behaviour in computers. Even if we enable autonomous systems to distinguish between more and less ethical alternatives, how can we be sure that they would choose right? We consider autonomous systems with a hybrid architecture in which the highest level of reasoning is executed by a rational (BDI) agent. For such a system, formal verification has been used successfully to prove that specific rules of behaviour are observed when making decisions. We propose a theoretical framework for ethical plan selection that can be formally verified. We implement a rational agent that incorporates a given ethical policy in its plan selection and show that we can formally verify that the agent chooses to execute, to the best of its beliefs, the most ethical available plan.

1. Introduction

Autonomous systems are increasingly required in various practical applications, including unmanned aircraft, driver-less cars, healthcare robots, manufacturing robots, etc. In all of these cases it is easy to imagine a situation where an autonomous system causes harm to people or property, as a result of an error in its engineering, or an unfortunate combination of circumstances. Therefore, if such autonomous systems are to operate within society, we must be able to trust that their behaviour complies with the legal, social, and ethical norms of that society. Determining the trustworthiness of technology in this respect is usually delegated to a regulatory body, such as the Federal Aviation Administration (for aircraft in the USA) or the Vehicle Certification Authority (for road vehicles in the UK). The process is known as *certification*, and is used to determine the safety and reliability of safety-critical technology, including aircraft, road vehicles, nuclear reactors, pharmaceuticals, etc.

For non-autonomous systems, such as cars or manned aircraft, it is assumed that the operator of the system will satisfy the ethical standards of society, *e.g.*, the pilot of a civilian aircraft does not intend to use the aircraft to commit murder, and will, if necessary, disregard legal restrictions for ethical reasons, *e.g.*, the pilot will disregard the Rules of the Air in order to preserve human life. These assumptions are an unavoidable result of the opacity of human behaviour; it is extremely difficult to predetermine the behaviour of a human being. However, autonomous systems are far more transparent, and can be engineered to meet requirements. Typically these requirements are technical ("an aircraft must be able to fly at 10,000 feet") or legal ("a car must

[*] Corresponding author.
E-mail addresses: L.A.Dennis@liverpool.ac.uk (L. Dennis), MFisher@liverpool.ac.uk (M. Fisher), Marija.Slavkovik@uib.no (M. Slavkovik), M.Webster@liverpool.ac.uk (M. Webster).

http://dx.doi.org/10.1016/j.robot.2015.11.012

have visible registration markings''), but in the case of autonomous systems some requirements may be ethical (*e.g.*, ''an autonomous unmanned aircraft will never choose to do something dangerous unless it has no other option''). Such ethical requirements may prove essential for an autonomous system to be certified by a regulatory body, since ethical autonomy is obviously desirable.

Machine ethics is an emerging discipline concerned with ensuring that the behaviour of machines towards humans and other machines they interact with is ethical [1]. It is an open question whether machines are, or will ever be, moral agents, *i.e.*, in possession of an innate ability to distinguish between right or wrong. However, it is necessary to enable them to adhere to our human understanding of morality, despite there exists no obvious or easy way to accomplish this [2–5].

If we assume that an autonomous system can be capable of moral agency, and possibly even be a better moral agent than a person, the goal of machine ethics is to enable machines to *reason ethically*. Notable works in this area are [6–11]. Within this subarea of machine ethics a lot of the questions traditionally studied in moral philosophy are reiterated but now from a computational perspective. The focus of research lies on automated extraction and identification of ethical guidelines for conduct, as well as on automated solving of ethical ambiguities and problems. These systems are often developed with the intention to be used to aid ethical decision-making by people.

If we assume that an autonomous system is *not* capable of moral agency, then the goal of machine ethics is to ensure that machines *behave ethically*. This is done by developing methods for ethically constraining the actions of machines [12]. Within this subarea of machine ethics, research focuses on identifying ethical principles that a system should not violate during its operation and developing methods for embedding consideration of these ethical principles in the decision-making process of the machine. Examples of work in this area are [13–15].

We are interested in representing and embedding consideration for ethical principles in the decision-making process of an autonomous system in a way that is amenable to certification. The work on ethically constraining actions of autonomous machines in [13–15] focuses on machines used in military operations and methods for stopping the autonomous machine from performing any action that is deemed unethical, but it does not consider circumstances where no ethical action is possible. Our focus in this paper is on civil applications. We propose a method for selecting among unethical actions, when no ethical action is possible, and for proving that a machine only behaves unethically, by choosing a minimally unethical course of action, if it has no ethical choice.

1.1. Formal verification

It appears increasingly the case, particularly in autonomous vehicles, that the autonomous control architecture is of *hybrid* form comprising discrete and continuous parts. Traditionally such systems have been engineered using the concept of a *hybrid automaton* (in which continuous aspects are encapsulated within a single state of an automaton while discrete jumps are represented as transitions between these states). However, as these systems have become more complex, combining discrete decision-making and continuous control in this way has created challenges for understandability and reuse of design and code.

Since we are particularly interested in the issue of decision-making, rather than control we have focused here on an alternative architecture, referred to as a *hybrid agent architecture* in which a distinguished agent is responsible for decision-making. This is motivated by evidence that hybrid automata based implementations scale poorly with the complexity of decision-making when compared to agent-based control [16,17]. A typical such architecture is shown in Fig. 1. The discrete part is often represented by a *rational agent* taking the high-level decisions, providing explanations of its choices, and invoking lower-level continuous procedures [18]. In this kind of hybrid autonomous system the continuous control and the higher-order decision-making components can be separated clearly. The lower-level procedures appear in non-autonomous systems as well, and are familiar to certification authorities. As such, we can focus analysis on the decisions the rational agent makes, given the beliefs and goals it has [19].

In an autonomous system we cannot show that an agent always does the right thing, but only that its actions are taken for the right reasons. Following this premise, *formal verification*, more precisely *model checking*, has been used in [20] for providing formal evidence for the certification of autonomous unmanned aircraft. Formal verification [21] involves proving or disproving that a system is compliant with a ''formally specified property'': a requirement specified in a mathematical language. Formal verification is an application of Formal Methods to the challenge of system verification. Model checking is a variety of formal verification in which *all* possible executions of a system are examined automatically based on a model of the real world. Model checking takes place relative to some requirement specified in a formal language [22].

In [19,20] *formal verification* is used to assess whether or not an autonomous system for an unmanned aircraft (UA) follows the specified ''Rules of the Air'' (ROA) that a pilot *should* follow [23]. The stated aim in these papers is to provide evidence that the autonomous system in control of an unmanned aircraft is safe and reliable, therefore providing supporting evidence for the potential certification of such an aircraft. The rationale behind using the Rules of the Air is that they provide a codified, statutory set of behaviours which human (and machine) pilots should satisfy. However, there are many circumstances that are not covered by the Rules of the Air. Indeed, the Rules of the Air are not intended to be exhaustive, but rather to provide a set of guidelines for pilot behaviour. It is anticipated that the Rules of the Air will be implemented by a skilled and experienced pilot whose responsibility is to ensure the safe passage of the aircraft through airspace (in this case, civil airspace). In those circumstances which are not covered by explicit Rules of the Air, it is the responsibility of the autonomous system in control of an unmanned aircraft to make sensible, rational, safe and *ethical* decisions at all times. So, while the formal verification of safe and legal decision-making has been covered in previous papers, we now focus on the formal verification of ethical decision-making within autonomous systems controlling autonomous aircraft.

1.2. Overview

This paper is organised as follows. In Section 2 we cover relevant background material on autonomous systems, machine ethics and verification. In Section 3 we outline our formal theoretical framework for the implementation and verification of ethically constrained behaviour in autonomous systems and also point to some relationships of our framework to deontic logic. In Section 4 we discuss our prototype implementation of this framework. In Section 5 we consider three simple examples of ethical reasoning implemented in our prototype, while, in Section 6, we present our conclusions and discuss further work.

2. Background

2.1. Agent architectures for autonomous systems

Webster et al. [19] discuss the analysis of an autonomous unmanned aircraft controller as a *hybrid system*, with an architecture such as the one given in Fig. 1. The rational agent-based