Verifiably Safe Decision-Making for Autonomous Systems

Yi Yang
imec-DistriNet, KU Leuven
Leuven, Belgium
yi.yang@kuleuven.be

ABSTRACT

Autonomous systems have the potential to significantly boost the productivity of our society. However, safety concerns are the primary impediment to the widespread use of autonomous systems. Safe decision-making for autonomous systems is a crucial step toward developing safe autonomous systems. My Ph.D. topic focuses on a formal approach to efficiently generating verifiable safe decision-making for autonomous systems. I have designed and implemented a three-stage formal approach to addressing the issue, and I have validated my approach with a real-world autonomous logistic system consisting of three autonomous mobile robots. This paper summarizes my current work and outlines my future work.

KEYWORDS

Autonomous System; Decision-Making; Formal Specification; Interpreter; Probabilistic Model Checking

1 INTRODUCTION

Autonomous systems have the potential to significantly boost the productivity of our society, as they can make their own decisions without external instructions. However, safety concerns have led many to express skepticism about autonomous systems such as self-driving cars. Therefore, providing safety assurance for autonomous systems is vital, albeit challenging. Safe autonomous decision-making is a crucial research topic for ensuring the safety of autonomous systems.

[15] provides a high-level overview of an approach for formally verifying the behaviors of autonomous systems. Building on the approach described in [15], [12] presents a verification methodology for the decision-making component in agent-based hybrid systems. [14] introduces an architectural framework for developing verifiable self-certifying autonomous systems. [16] analyzes the key aspects to develop a framework for the certification of reliable autonomous systems. [10] presents a framework for verifiable autonomous decisions and its applications to assessing a range of properties of autonomous systems.

My research builds upon the assumptions and definitions established in the aforementioned research work [10, 12, 14–16]. Specifically, I focus on the design and analysis of a safe agent-based decision-making component that serves as a high-level discrete controller for an autonomous system. This component operates independently but collaboratively with the low-level continuous controller of an autonomous system [12]. Given the inherent unpredictability of the real world, it is impossible to guarantee that any system will always behave safely [10]. Consequently, I consider an autonomous decision-making component safe if it avoids deliberately pursuing unsafe behaviors based on its beliefs and goals [10]. Furthermore, I use formal verification to provide safety assurance for the decision-making component.

2 A THREE-STAGE FORMAL APPROACH

To facilitate verifiable safe decision-making for autonomous systems, I have devised a three-stage formal approach consisting of formal specification, safe decision generation, and PCTL model checking. This formal approach is made possible through the utilization of four key components: a specification language named vGOAL that specifies autonomous decision-making mechanisms; an interpreter for vGOAL that automates safe decision-making generation; a translator that translates a given vGOAL specification to a PRISM model; and a PCTL model checker such as Storm [9] or PRISM [22] that verifies the soundness of the given vGOAL specification. My contributions are vGOAL, the interpreter of vGOAL, and the translator of vGOAL. The approach has been validated by a real-world autonomous logistic system consisting of three autonomous mobile robots. Three demos can be accessed at [26]: one demo for an error-free run, one demo for a run including a non-fatal error, and one demo for a run including a fatal error. The following briefly explains the key aspects of each stage.

2.1 Formal Specification: vGOAL

The formal specification requires a specification language that is expressive to specify autonomous decision-making mechanisms and suitable for formal verification. Additionally, it is advantageous to have a compatible interface with a widely used development framework for robotic applications.

Agent programming languages (APLs), including AgentSpeak [3], Jason [4], Gwendo [11], and GOAL [18], have been extensively researched for programming autonomous agents for decades, rendering them well-suited for specifying autonomous decision-making mechanisms. GOAL shares many features with Belief-Desire-Intention (BDI) APLs, such as beliefs and goals, but it is primarily a rule-based APL [5]. Despite the potential benefits of APLs in the development of autonomous robotic systems, their research has not been widely used in the field. Integration with the Robot Operating System (ROS) may expand their applications to robotics, as ROS has become the de facto standard for developing robotic applications.
**vGOAL** is motivated by three primary considerations, which is a GOAL-based specification language that focuses exclusively on the internal logic reasoning mechanism of GOAL. First, GOAL is highly suitable for specifying autonomous decision-making, but many of its specifications are irrelevant to this domain, such as environment specifications. Second, the intrinsic logic-based nature of GOAL makes it highly suitable for formal verification. Third, GOAL has no build-in interface to ROS, which limits its applicability in robotic applications. Therefore, vGOAL can leverage the strengths of GOAL, formal verification, and ROS.

### 2.2 Safe Decision Generation: Interpreter

The agent-based decision-making component is implemented as the interpreter for vGOAL, which is integrated into ROS via rosbridge [7]. The motivation and the initial design of this stage were described in [27].

Figure 1 demonstrates how the interpreter generates safe decisions through interaction with ROS. ROS keeps sending real-time sensor information to the interpreter on a regular basis, e.g., every 100 milliseconds. The interpreter has three main components: a data processing component, a decision-making generation component, and a safety checking component. The interpreter generates decisions based on real-time information. Each generated decision will be checked if it violates any safety requirements. Only safe decisions can be sent to agents.

### 2.3 PCTL Model Checking: Translator

A vGOAL specification is considered sound when there is at least one feasible and safe plan available to achieve all goals. Formal verification is a compelling method for verifying the soundness of vGOAL specifications. Owing to the automated verification process, model checking is the most successful and influential verification method in verifying APLs, such as AgentSpeak(L) [2], Gwendolen, GOAL, SAAPL [24], and ORWELL [8] [13] [23] [28]. My approach encodes each state with its beliefs, allowing no additional computation for safety checking. This state encoding is easily expressible in the PRISM language, unlike in CTL model checkers such as SPIN [19] and NuSMV [6]. A finite transition system can be converted into a semantically equivalent finite discrete-time Markov chain (DTMC) except for transition probabilities. Moreover, the verification result of qualitative properties including safety and liveness properties in a finite DTMC is irrelevant to the transition probabilities [1]. Therefore, PCTL model checking is chosen to verify the soundness of vGOAL specifications.

Figure 2 illustrates the workflow of the automated PCTL model checking process. A translator was developed to convert a vGOAL specification into a PRISM model, with two components: transition system generation and PRISM model encoding. The translator generates the operational-semantically transition system from the vGOAL specification and encodes it as a semantically equivalent PRISM model except for transition probabilities. The translator and a PCTL model checker (Storm or PRISM) will automatically generate a PCTL model checking analysis from a vGOAL specification.

### 3 Future Work

My future work includes two main directions. First, I plan to employ formal verification techniques, such as program verification, to prove the correctness of the implementation of the vGOAL interpreter. Second, I aim to investigate how a more sophisticated, safe, and intelligent motion planning component can be integrated into the agent-based decision-making module. Recent safe shielding techniques [17, 20, 21, 25] enable reinforcement learning-based control of autonomous systems in continuous state spaces while ensuring safety. It would be valuable to integrate the recent shielding techniques into the described agent-based decision-making component to provide safety assurance for autonomous systems.
ACKNOWLEDGEMENTS

This research is partially funded by the Research Fund KU Leuven.

REFERENCES