Reasonableness Monitors

Leilani H. Gilpin
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street
Cambridge, MA 02139
{ lgilpin } @ mit.edu

Abstract

As we move towards autonomous machines responsible for making decisions previously entrusted to humans, there is an immediate need for machines to be able to explain their behavior and defend the reasonableness of their actions. To implement this vision, each part of a machine should be aware of the behavior of the other parts that they cooperate with. Each part must be able to explain the observed behavior of those neighbors in the context of the shared goal for the local community. If such an explanation cannot be made, it is evidence that either a part has failed (or was subverted) or the communication has failed. The development of reasonableness monitors is work towards generalizing that vision, with the intention of developing a system-construction methodology that enhances both robustness and security, at runtime (not static compile time), by dynamic checking and explaining of the behaviors of parts and subsystems for reasonableness in context.

Introduction

An important problem of complex autonomous systems is that they cannot provide insight into their behaviors and thought processes. This work on reasonableness monitoring is a first step towards developing the methodologies and technologies to support building robust, articulate systems. Such systems will be introspective and able to explain their behaviors. They will be able to use explanations to dynamically detect, and mitigate, anomalous behaviors.

Reasonableness monitors are implemented as two types of interfaces around the subsystems of a complex machine. Local monitors dynamically check the behavior of a specific subsystem, and non-local reasonableness monitors monitor committees of subsystems for plausibility (reasonableness) in context. Reasonableness monitors are able to build an explanation of a problem (or a reasonable state) by examining the premises supporting the observation of a contradiction (or consistency).

I present preliminary results on applying a local reasonableness monitoring system to machine perception. The key contribution here is determining the reasonableness of a perception-derived scene description with the careful use of dependencies and dependency analysis. I also present preliminary technical contributions and a methodology towards monitoring non-local inconsistencies. This methodology is inspired by the structure of successful human organizations, where tasks are accomplished by committees of multiple people. Committees are able to survive bad work by any single member, because members of the committee observe each other's work and can jointly decide on actions to correct bad work or remove misbehaving members. The goal of reasonableness monitoring is to apply this philosophy of committees to subsystems of machines, with the aim of producing more robust and secure systems for safety-critical or mission-critical tasks.

Methods

Each reasonableness monitor has its own knowledge base: a set of behaviors that are considered to be reasonable. For monitoring machine perception, ConceptNet 5 (Speer and Havasi 2013) is used as a knowledge base of reasonableness. Monitors search through the knowledge base for premises and generate explanations of inconsistencies. The premises are used as evidence for explaining inconsistent (or consistent) information.

These methods are greatly influenced by work on knowledge representation (Fahlman 1979), analogical chaining (Blass and Forbus 2017), and monitoring systems for planning (Veloso, Pollack, and Cox 1998). Explanatory systems have also been applied to multi-agent domains (Molineaux and Aha 2015) and story understanding (Winston and Holmes 2017).

Preliminary Results

Take the observed perception of "a mailbox crossing the street." This is clearly unreasonable since mailboxes are heavy, inanimate objects found at the sidewalk of a street that do not move. Reasonableness monitors can detect and explain this unreasonable perception.

```
input: "A mailbox crossing the street"
parsed as: (mailbox, cross, street)

This perception is UNREASONABLE,
  using data from ConceptNet5.
```

```
REASONING:
A mailbox is a type of box typically
found near a sidewalk.

Mailboxes cannot cross a street
because mailboxes are objects that
do not move on their own.
```

Reasonableness monitors can also explain and detect consistent information. This monitor can explain that the perception of "a mailbox contains papers" is reasonable.

```
input: "A mailbox contains papers"
parsed as: (mailbox, contain, papers)

This perception is REASONABLE,
  using data from ConceptNet5.

REASONING:
A mailbox is used for receiving
letters.  And a letter is a sub type
or specific instance of a piece
of paper.

So mailboxes can reasonably contain
papers.
```

Future Work

Current work is focused on creating the system design to monitor non-local inconsistencies. For instance, the monitor may be provided with additional alternative premises explicitly (e.g a mailbox in a hurricane can move). These sorts of premises can be manually inserted into the monitor and cause previously reasonable premises to lose support.

Longer term contributions are focused on learning shared premises (and ranking their applicability of support) between subsystems. This will rely on the structure of committees of multiple subsystems. When a contradiction is exposed, if more than one subsystem provides an acceptable explanation, then the committee decides which explanation is most appropriate. If the explanations are all logical, then the premises that support the alternative propositions are inspected by the committee. The committee decides which premises should prevail and thereby which conclusion should be accepted.

Non-local Monitors

I extended the monitoring system to be able to reconcile a previously inconsistent system state. In this proof-of-concept, premises can be manually added. Returning to the perception of "A mailbox crossing the street", if the premise: `(mailbox in hurricane, move, plausible)` is added, this premise takes precedent, and the `(mailbox, moves, false)` premise is removed. The system is able to explain that the perception is reasonable with this added premise:

```
added premise :
(mailbox in hurricane, move, plausible)

REASONING:
Typically, mailboxes cannot cross a
street because mailboxes are objects
that do not move on their own.

But during a hurricane, it's plausible
for a mailbox to move.  Therefore,
it is reasonable for a mailbox to
cross the street during a hurricane.
```

Learning Shared Premises

The question remains how to learn premises that are shared among parts. When monitoring a community of parts working together, the set of local premises that support the behavior may not be enough to produce a coherent explanation. Further, these monitors will need to determine which premises take precedent when there are inconsistencies. Currently, for our proof-of-concept, the most recently added premise takes precedent. However when neighboring monitors are alerted of inconsistencies, the most supported premise may not obvious. The idea is to diligently use dependencies, dependency tracking, and dependency analysis. Once the offending premise is removed, we are left with a consistent worldview. However, not all non-local consistencies will be this simple to explain and detect.

Returning to the "mailbox crossing the street' example, I propose that the alternative support will be learned with respect to a more general knowledge base of reasonableness (e.g. high winds can cause immobile heavy objects to move) and deduction (high winds can move heavy objects and a mailbox is a heavy object, therefore high winds can move a mailbox. Since a hurricane is characterized by high winds, it is plausible for a mailbox to be perceived to walk across the street during a hurricane). Both approaches arrive at the same conclusion and produce an equally sufficient explanation. The methods and computational techniques for learning this type of shared knowledge and premises between parts are left to future work.

References

Blass, J. A., and Forbus, K. D. 2017. Analogical chaining with natural language instruction for commonsense reasoning. In *AAAI*, 4357–4363.

Fahlman, S. E. 1979. *NETL, a system for representing and using real-world knowledge*. MIT press.

Molineaux, M., and Aha, D. W. 2015. Continuous explanation generation in a multi-agent domain. Technical report, NAVAL RESEARCH LAB WASHINGTON DC.

Speer, R., and Havasi, C. 2013. Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*. Springer. 161–176.

Veloso, M. M.; Pollack, M. E.; and Cox, M. T. 1998. Rationale-based monitoring for planning in dynamic environments.

Winston, P. H., and Holmes, D. 2017. The genesis manifesto: Story understanding and human intelligence.