An Answer Set Programming Framework for Reasoning about Agents' Beliefs and Truthfulness of Statements

Marcello Balduccini, Michael Gelfond, and
Enrico Pontelli and Tran Cao Son


Computer Science Department, New Mexico State University, Las Cruces, NM


Computer Science Department, Texas Tech University, Lubbock, TX


Department of Decision & System Sciences, Saint Joseph's University, Philadelphia, PA

KRR 2020

Outline

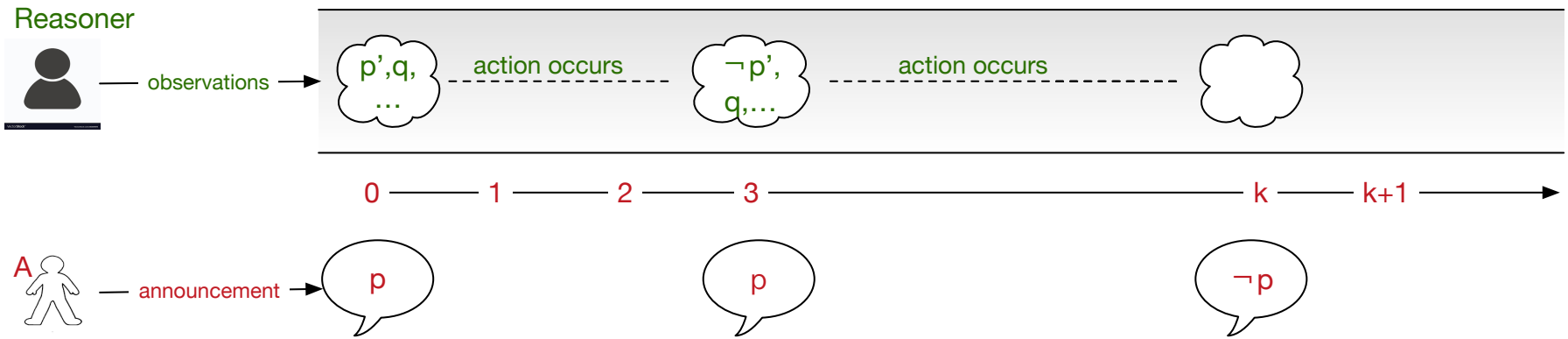
Example

Time step	Information
0	John says that his family is poor (*poor*).
1	We observe that John attends an expensive college (*expensive_college*).
2	We learn that John has a full scholarship because of his financial hardship (*has_scholarship*).

Is John's statement about his family's financial status truthful?

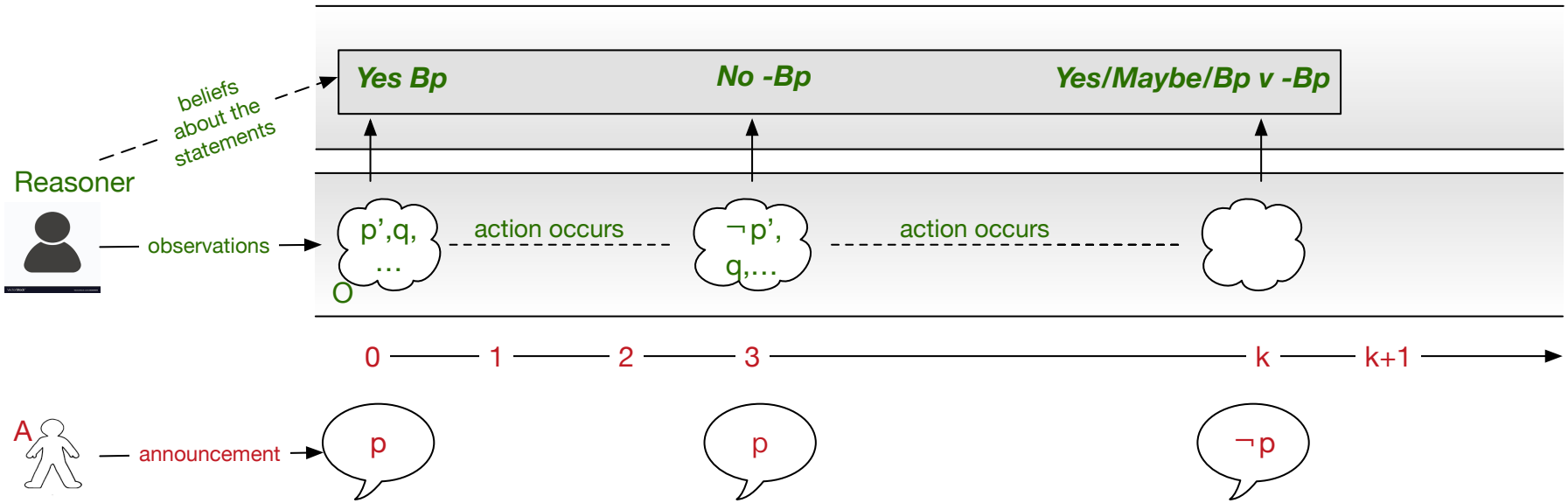
- at time 0? Yes (there is nothing that proves otherwise.)
- at time 1? No (by default, attending expensive college requires lots of money.) [d_1]
- at time 2? Yes (by default, financial hardship scholarship is only for low income families.) [d_2 and d_2 is **more preferred than** d_1]

In everyday life, agents observe the world, receive information from others, and make judgment about the truthfulness of information provided by others. How do they reach their conclusion?



Questions: (by the reasoner) does A tell the truth?

In everyday life, agents observe the world, receive information from others, and make judgment about the truthfulness of information provided by others. How do they reach their conclusion?



Questions:

1. Can the statement about property *p* made by agent *A* believed to be true?
2. Which statements of other agents does the reasoner believe to be true? false?

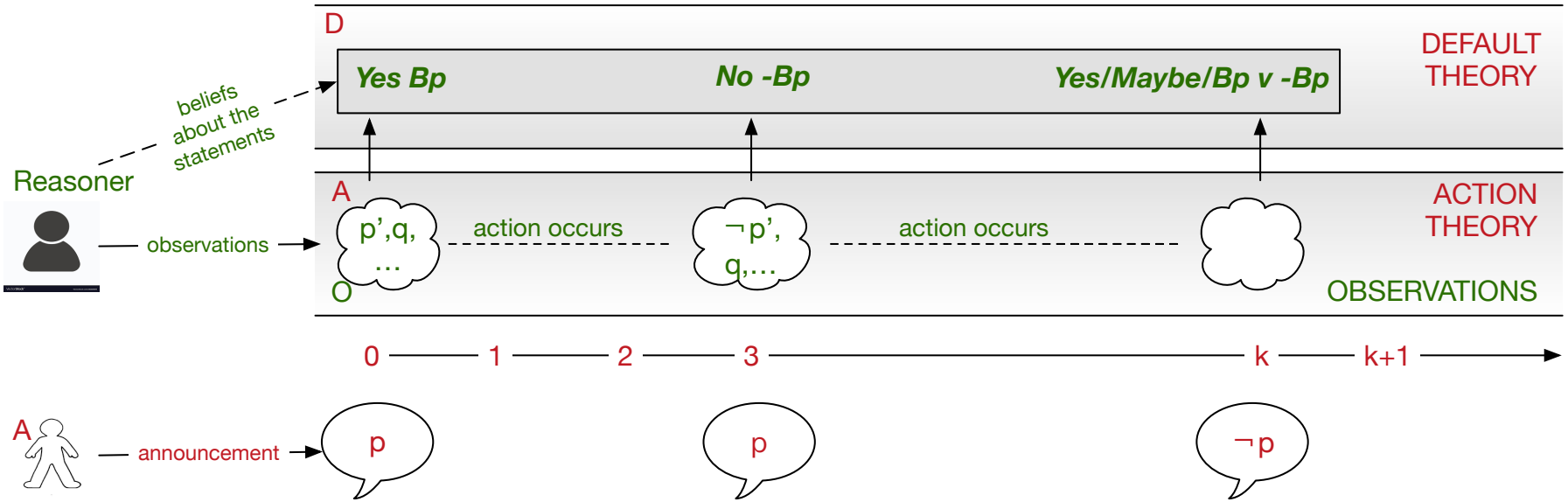
Contributions

1. An abstract model for representing and reasoning about
   1. the evolution of an agent's beliefs over time; and
   2. truthfulness of third-party statements;
2. A concrete realization of the model using Answer Set Programming; and
3. A demonstration of the proposed framework on an important problem in the area of cyber security.

Outline

Assumptions (about the reasoner)

- optimistic in nature: an agent believes a statement made by another agent unless there is a reason to believe otherwise!
- commonsense reasoner: an agent uses default reasoning to draw conclusions about truthfulness of statements made by other agents.
- observant: an agent observes other agents and will use her observations in draw conclusions.
- active: an agent executes her own actions to change the world.



Solution

- Represent the knowledge base and default rules of the agent by a default theory *Def*.
- Represent actions and their effects by an action theory *A*.
- Represent observations by a set of facts *O*.
- Separate beliefs of the reasoner and the properties of the world.
- Develop method for evaluation of statements of agents (define $Def \cup A \cup O \models p[t]$ for p is believed to be true at time t).

- Given: $T = (O_a, O_f, Act, Def)$ and $\models_A$ and $\models_D$.
  - Action theory: Act in a suitable logic A that defines $\models_A$

$$Act \cup O_a \cup I \models_A p \textbf{ after } [a_0, \ldots, a_n]$$

    p is true after the execution of $[a_0, \ldots, a_n]$ from the state I.
  - Default theory: Def defines $\models_D$, $Def \cup O \models b$
    b is true w.r.t. the set of observations O and the theory Def.
- Question: Is b true (false) at a certain time step t, denoted by $p[t]$, given T?

$$T \models b[t] \Leftrightarrow \langle W[t], Def \rangle \models_D b$$

Steps in determining $b[t]$:

- $W[t]$: model of the world at the time step t from Act, O_a, and O_f (using $\models_A$); and
- Determine whether b is true given Def and $W[t]$ (using $\models_D$).

Outline

Representation of $T = (O_a, O_w, Act, Def)$

- Observations:
  - $obs(p, s)$: proposition p is observed at time step s.
  - $occ(a, s)$: action a occurred at time step s.
- Action theory: high-level action language [1].
- Default theory: default theory with preferences [2].

Encoding of T by $\Pi(T)$, an answer set program with rules for

1. reasoning about actions and observations, defining $holds(f, t)$: f is true at step t;
2. reasoning about defaults, defining $believes(b, t)$: b is believed to be true at step t.

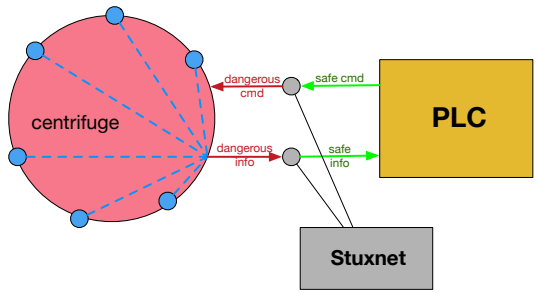
Entailment between T and $stm(b, s)$, $t \geq s$

- $T \models +stm(b, s)@t$ if $\Pi(T) \models believes(b, t)$;
- $T \models -stm(b, s)@t$ if $\Pi(T) \models believes(\overline{b}, t)$
- $T \not\models \pm stm(b, s)@t$ if $T \not\models +stm(b, s)@t$ and $T \not\models -stm(b, s)@t$.

Outline

- **Stuxnet**: manipulates the centrifuge and falsely informs controller (PLC)
- Could be prevented if additional information is collected!

default(d_1^m, hot_room, [alert]).
default(d_2^m, ¬hot_room, [cold_weather]).
prefer(d_2^m,d_1^m, [ ]).
rule(r_1^m,cold_weather,[winter]).
default(d_3^m,overheat,[alert, ¬hot_room]).



Outline

- Propose a general declarative framework for representing and reasoning about truthfulness of agents from
  - observations about the state of the world;
  - knowledge about the actions of the agents; and
  - normal behavior of agents.
- Develop an implementation in ASP.
- Present an application of the proposed framework in detecting man-in-the-middle attacks targeting computer and cyber-physical systems.

Application of the system in real-world domain:

1. reasoning about reputation of agents (can we trust agent *A*?)
2. diagnostic reasoning about statements by agents (what should be done to confirm/reject a statement by agent *A*?)
3. integrating with a model of trust (should I trust information *X* from agent *A* given that I trust him only 60% of the time?)

Thank you for your attention.

References

M. Gelfond and V. Lifschitz.
Action Languages.
Electronic Transactions on Artificial Intelligence, 3(6), 1998.

M. Gelfond and Tran Cao Son.
Prioritized default theory.
In *Selected Papers from the Workshop on Logic Programming and Knowledge Representation 1997*, pages 164–223. Springer Verlag, LNAI 1471, 1998.