Reasoning about Cognitive Trust in Practice: Robotic Waiter Example

Maciej Olejnik, Marta Kwiatkowska
Department of Computer Science
UNIVERSITY OF OXFORD

Motivation

- Autonomous robots entering our society – can we trust them?
- Driverless cars will need to share roads with human drivers.
- Fatal Tesla crash – driver's overtrust caused the accident.

Trust Formulations

- Subjective evaluation of a trustor on a trustee, relative to a task.
- Informed by past experience, social norms and preferences.
- Cognitive trust captures the human notion of trust.
  - Relative to a set of goals.
  - Expressed in terms of beliefs.

Aim of This Work

- Put forward a framework for reasoning about cognitive processes and trust in stochastic multi-agent systems.
- Propose a logic to reason about mental attitudes and express trust-related concepts such as competence, disposition or dependence.
- Develop tools for verifying cognitive and trust properties specified using the aforementioned formalism.

Framework

An Autonomous Stochastic Multi-Agent System (ASMAS) consists of:

- Ags – set of agents
- S – set of states
- $\{Act_A\}_{A \in Ags}$ – set of actions for each agent
- $T : S \times Act \to \mathcal{D}(S)$ – transition function
- $L : S \to \mathcal{P}(AP)$ – labelling function
- $\{O_A\}_{A \in Ags}$ – set of observations for each agent
- $\{obs_A\}_{A \in Ags}$ – observation function for each agent (where $obs_A : S \to O_A$)
- $\{Goal_A, Int_A\}_{A \in Ags}$ – goals and intentions of each agent
  - Goals are static, abstract mental attitudes
  - Intentions are dynamic plans of execution (strategies)
  - Agents change goals or intentions via cognitive transitions
- $\{gp_{A,B}, ip_{A,B} \mid B \in Ags\}_{A \in Ags}$ – set of preference functions for each agent
  - They represent agents' prior knowledge of each other
- $\{\pi^g_A, \pi^i_A\}_{A \in Ags}$ – cognitive strategy of each agent.

Logic

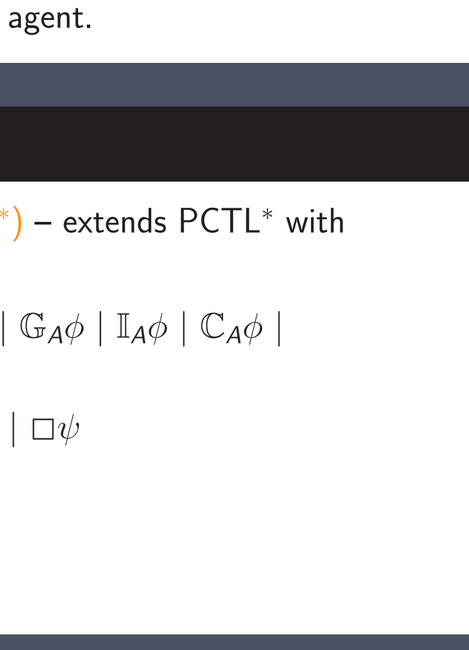
Probabilistic Rational Temporal Logic (PRTL*) – extends PCTL* with cognitive and trust operators.

$$\phi ::= p \mid \neg\phi \mid \phi \vee \phi \mid \forall\psi \mid \mathrm{P}^{\bowtie q}\psi \mid \mathbb{G}_A\phi \mid \mathbb{I}_A\phi \mid \mathbb{C}_A\phi \mid \mathbb{B}^{\bowtie q}_A\psi \mid \mathbb{CT}^{\bowtie q}_{A,B}\psi \mid \mathbb{DT}^{\bowtie q}_{A,B}\psi$$
$$\psi ::= \phi \mid \neg\psi \mid \psi \vee \psi \mid \bigcirc\psi \mid \psi \mathrm{U} \psi \mid \Box\psi$$

- $\mathbb{G}_A$, $\mathbb{I}_A$, $\mathbb{C}_A$ are cognitive operators
- $\mathbb{B}^{\bowtie q}_A$ is a belief operator
- $\mathbb{CT}^{\bowtie q}_{A,B}$, $\mathbb{DT}^{\bowtie q}_{A,B}$ are trust operators

Robotic Waiter Example

Our example is set in a restaurant in which customers are served by robotic waiters. We assume that robots take orders from customers (we distinguish cheap and expensive orders) and assign high, medium or low priority to each. We also assume a mechanism that incentivises waiters to assign low priorities to orders. Depending on their priority, orders take different amount of time to be prepared and as a result will be delivered (probabilistically, see Figure 4) late or on time. Finally, the customer may optionally leave a tip, which can be high or low.



Modelling

Construct an ASMAS model of Robotic Waiter Example:

- $Ags = \{\text{Rob}, \text{Charlie}\}$ – Rob is the waiter, Charlie the customer
- $Act_{Rob} = \{\text{low}, \text{med}, \text{high}\}$
- $Act_{Charlie} = \{\text{cheap}, \text{expensive}, \text{none}, \text{low}, \text{high}\}$
- T represented graphically below – agents settle their goals $\to$ Charlie orders $\to$ Rob assigns priority to the order $\to$ Charlie gives (or not) a tip
- $AP = \{onTime, late\}$, L as in Figure 4.
- Agents cannot observe other agent's goals
- $Goal_{Rob} = \{\text{fair}, \text{greedy}, \text{opportunistic}, \text{strategic}\}$
- $Goal_{Charlie} = \{\text{mean}, \text{generous}, \text{extravagant}, \text{prudent}\}$
- $gp_{Charlie,Rob}(s_0) = \langle \text{fair} \mapsto 0.7, \text{greedy} \mapsto 0.3 \rangle$
- $gp_{Rob,Charlie}(s_1) = \langle \text{extravagant} \mapsto 0.5, \text{prudent} \mapsto 0.5 \rangle$

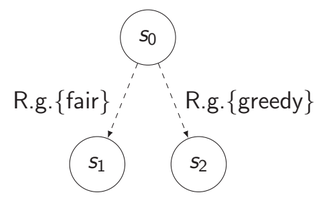


Figure 1: Rob's goal change

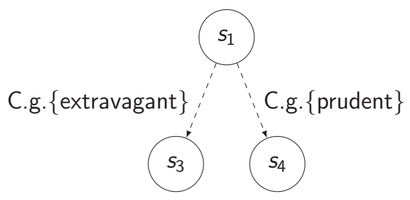


Figure 2: Charlie's goal change

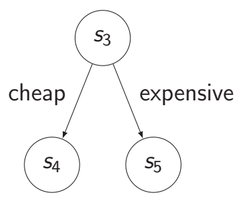


Figure 3: Charlie's order

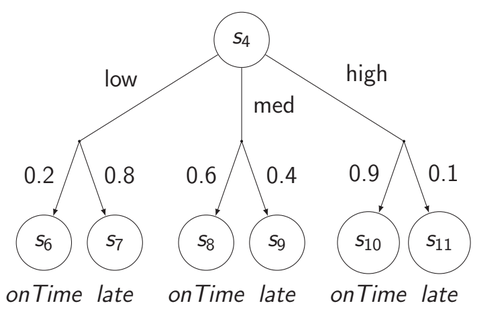


Figure 4: Meal delivery probabilities

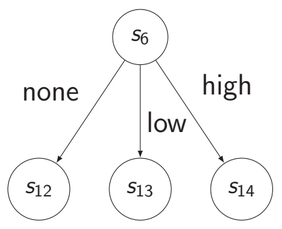


Figure 5: Charlie's tip

order / strategy	cheap	expensive
$\sigma_{extravagant}$	0.2	0.8
$\sigma_{prudent}$	0.9	0.1

Table 1: Order strategy

Expressing Belief

Belief – fundamental notion in ASMAS. We give few example formulas and evaluate them assuming the above set up:

- $\mathbb{B}^{=?}_{Charlie} fair_{Rob}$ – queries Charlie's belief that Rob is fair. Evaluates to 0.7.
- $\mathbb{B}^{\geq 0.2}_{Charlie} \Diamond onTime$ – expresses that Charlie's belief that his order will be delivered on time is at least 20%. This formula is true.
- $\mathbb{B}^{=?}_{Rob} \bigcirc expensive_{Rob}$ – queries Rob's belief that Charlie will order an expensive meal. Evaluates to 0.45.

Cognitive and Trust Formulas

Below are few examples of more complex formulas involving trust and cognitive operators, which demonstrate the expressive power of PRTL*:

- $\mathbb{G}_{Rob}\mathrm{P}^{\leq 0.9}\Diamond onTime$ – "Regardless of Rob changing his goals, the probability of the meal arriving on time is no greater than 90%".
- $\bar{\mathbb{I}}_{Charlie}\mathbb{G}_{Rob}\mathrm{P}^{\geq 0.5}\Diamond onTime$ – "Charlie can change his intention, so that, regardless of Rob changing his goals, the probability of the meal arriving on time is at least 50%".
- $\mathbb{CT}^{\geq 0.7}_{Charlie,Rob}\Box(expMeal \to \mathrm{P}^{\geq 0.9}\Diamond onTime)$ – "Charlie's trust in Rob's capability of ensuring that the meal will arrive on time with probability at least 90% is greater or equal to 70%".
- $\mathbb{B}^{\geq 0.8}_{Rob}\mathbb{DT}^{\geq 0.7}_{Charlie,Rob} greedy_{Rob}$ – "Rob's belief that Charlie has at least 70% trust in him being greedy is at least 80%".

Future Work

- Existing model checkers, such as PRISM (www.prismmodelchecker.org), helpful, but do not support ASMAS semantics.
- Need a tool for verifying ASMAS automatically.

Acknowledgments

The authors are supported by EPSRC Mobile Autonomy Programme Grant EP/M019918/1.