Reasoning about Cognitive Trust in Practice: Robotic Waiter Example

Maciej Olejnik, Marta Kwiatkowska
Department of Computer Science
University of Oxford

***Abstract*—Technological advances have brought about machines capable of beating humans in chess or, more recently, in an even more complex game of Go. While those milestones might seem abstract to many, self-driving cars or medical robots provide an example which should not be ignored. Given that human lives are at stake, it is important to consider whether (or to what degree) we can trust autonomous systems. As a step towards addressing that question, [5] presented an automated verification framework for reasoning about cognitive processes and trust in stochastic multi-agent systems. The formalism captures human notion of trust, defined as a subjective evaluation of agent A on agent B's ability to complete a task, which may lead to a decision of A to rely on B. A probabilistic rational temporal logic PRTL* was introduced, which extends PCTL* with novel operators that let one reason about mental attitudes and express trust-related concepts such as competence, disposition or dependence. In this work, we illustrate usefulness of the framework on the *Robotic Waiter Example*, which investigates human-robot interaction in a real-world setting.**

I. Introduction

Recent years have seen rapid progress of autonomous robotics, with self-driving cars, home assistive robots or unmanned aerial vehicles entering the fabric of our society. An important question arising in that context is whether we are safe in presence of this new technology. A recent fatal collision involving a Tesla car in autonomous mode [6] shows the potential dangers of deploying autonomous robots and vehicles on a wide scale. Importantly, the driver's over-reliance on the car's software has been identified as the main cause of the crash, thereby providing us with motivation for studying cognitive trust between humans and robots.

In most general terms, trust is understood as a *subjective evaluation of a trustor on a trustee about something in particular*, e.g., completion of a task [4]. In this work, we focus on relationships between humans and autonomous systems and are primarily interested in *cognitive* trust, which reflects aspects such as human motivation, goals and intentions, as well as the social context. We follow [1] and view trust as a complex *mental attitude* that is relative to a set of goals and expressed in terms of beliefs, which in turn influence decisions about agent's future behaviour.

The framework for reasoning about cognitive trust comprises a model called Autonomous Stochastic Multi-Agent System (ASMAS), which consists of a set of *agents*, each equipped with a set of local *actions*, interacting within an environment. It evolves by transitioning between *states* according to the *transition function*, where each transition is caused by agents selecting their local actions independently and executing them simultaneously. A distinctive aspect of ASMAS is that it combines the *temporal* dimension, i.e., actions of agents in a physical space described above, with the *cognitive* dimension, which concerns agents' cognitive (i.e., mental) changes. In ASMAS, the two dimensions coexist, and each transition of the system belongs to one of them.

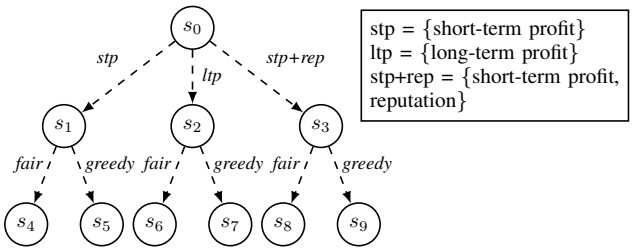

Fig. 1. Cognitive transitions of robotic waiter

II. Modelling the Robotic Waiter Example

We present the details of our framework with help of a simple example involving interaction between humans and robots, which we refer to as *Robotic Waiter Example*.

A. Setting

Our scenario, inspired by Rong Heng Seafood Restaurant in Singapore [2], is set in a restaurant, in which customers are served by autonomous robots. Each guest of the restaurant may order an expensive meal or a cheap meal, which then enters the queue of orders, according to the priority assigned to it by the waiter. Depending on its position in the queue, the dish might arrive on time, in which case the waiter receives a bonus proportional to the value of the meal, or late, in which case the waiter receives no bonus. Some waiters are fair, and assign the same priority to all the orders they take, but others are greedy and assign higher priority to more expensive dishes, in order to maximise their gratification.

We aim to reason about the degree of trust a customer has in the robotic waiter to deliver their meal on time, and how robot's mental state, i.e., being fair or greedy, affects their behaviour.

B. The Model

We now give a more detailed overview of the main components of ASMAS and develop a model of the Robotic Waiter

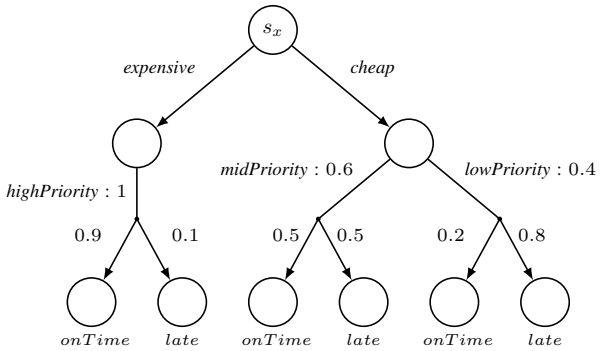

Fig. 2. Temporal dimension corresponding to greedy waiter, $x \in \{5, 7, 9\}$

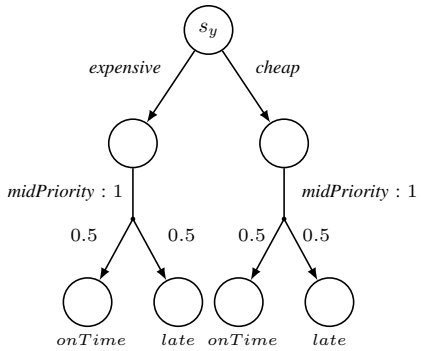

Fig. 3. Temporal dimension corresponding to fair waiter, $y \in \{4, 6, 8\}$

Example step by step.

1) Cognitive state: We represent agents' mental attitudes by equipping each of them with a set of possible goals and a set of possible intentions. In our model of Robotic Waiter Example, there are two agents: Charlie (the customer) and Rob (robotic waiter). For simplicity, we only consider the Rob's cognitive processes. His set of possible goals is $Goal_{Rob} = \{$*short-term profit*, *long-term profit*, *reputation*$\}$, and the set of possible intentions is $Int_{Rob} = \{$*fair*, *greedy*$\}$. In general, goals are high-level, abstract entities and represent the desires of an agent, while intentions are more concrete and often realised as means of achieving a given set of goals.

While sets $Goal_A$ and Int_A of agent A are static, in a sense that they do not change throughout the execution of a system, the dynamic nature of agents' mental attitudes is captured by their *cognitive state*, consisting of a set of goals (a subset of $Goal_A$) and a single intention (an element of Int_A). Figure 1 illustrates possible cognitive transitions (i.e., changes of cognitive state) of Rob.

Finally, we note that, since our work focuses on agents' mental processes, we assume that cognitive state determines agents' behaviour in temporal dimension. In the Robotic Waiter Example, this is realised by associating each intention with an action strategy. We now extend our model to include temporal actions of Charlie and Rob. Following Rob's cognitive transitions, Charlie performs their temporal action, one of *expensive*, *cheap*, corresponding to ordering expensive or cheap meal, respectively. Afterwards, Rob assigns priority to the order, by performing one of three actions: *highPriority*, *midPriority* or *lowPriority*, each of which results in a probabilistic transition to one of two states, depending on whether the meal arrives on time or late. Figure 2 illustrates the unfolding of the system when Rob is greedy, i.e., from states s_5, s_7, s_9. His action strategy is represented in the figure by the probability value next to each of his actions. The probability values in the figure can be statistically inferred from past data. Similarly, Figure 3 shows how the system evolves when Rob is fair.

2) Cognitive mechanism: Note that there are subsets of $Goal_{Rob}$ which are not included in Figure 1. Some of them, such as $\{$*short-term profit*, *long-term profit*$\}$, are inconsistent, while others, e.g., $\{$*reputation*$\}$, are omitted for simplicity. Formally, possible goal or intention changes of a given agent A are specified by the *legal goal function* $goal_A$ and the *legal intention function* int_A. We refer to a pair $\langle goal_A, int_A \rangle$ as the *cognitive mechanism* of agent A. Using abbreviations from Figure 1, the cognitive mechanism of Rob is:

$$goal_{Rob}(s_0) = \{\textit{stp}, \textit{ltp}, \textit{stp+rep}\},$$
$$int_{Rob}(s_x) = \{\textit{greedy}, \textit{fair}\},$$

where $x \in \{1, 2, 3\}$.

3) Cognitive strategy: Analogously to action strategies in temporal dimension, agents use cognitive strategies to determine what mental changes to perform for a given execution history. For example, the following intention strategy:

$$\pi^i_{Rob}(s_0 s_1) = \langle \textit{greedy} \mapsto 1 \rangle,$$
$$\pi^i_{Rob}(s_0 s_2) = \langle \textit{fair} \mapsto 1 \rangle,$$
$$\pi^i_{Rob}(s_0 s_3) = \langle \textit{greedy} \mapsto 1/2, \textit{fair} \mapsto 1/2 \rangle,$$

indicates that Rob is greedy when it aims to profit in the short term, fair, when it aims to profit in the long term, and greedy or fair, with equal probabilities, when it aims to profit in the short term and retain its reputation.

4) Partial observability: An important notion, inherent in the semantics of ASMAS, is partial observability. It arises due to the nature of cognitive state of an agent, which is generally not observable to other agents. For instance, in the Robotic Waiter Example, Charlie does not know the goals and intention of Rob. Formally, we say that Charlie's observation cannot distinguish states s_1, s_2 and s_3. However, we assume that agents can observe the number of states during a given execution (in other words, they observe that transition happened, but they may not know what that transition was). Therefore, even though Charlie cannot tell states $s_4, \ldots, s_9$ apart, he can differentiate them from states s_1, s_2, s_3. On the other hand, Rob can observe its own cognitive changes and so its observations are unique for all states.

5) Preference functions: In order to reason formally about a given ASMAS, one must quantify the likelihood of different paths in the system, which can be achieved by defining a probability space on the set of all paths. In ASMAS,

nondeterminism in temporal dimension is resolved by our assumption that cognitive state induces an action strategy for each agent, as we saw for the waiter in our example. In cognitive dimension, nondeterminism is resolved by each agent's *preference functions*, which represent prior knowledge one agent has about another.

For example, Charlie might have experienced that, when he is served by Rob, expensive dishes usually arrive on time, but cheap meals are frequently late. Charlie knows that waiters aiming for short-term profit are always greedy (i.e., his intention preference function over Rob – or any other robotic waiter – on state s_1 gives $ip_{Charlie,R1}(s_1) = \langle \textit{greedy} \mapsto 1 \rangle$). Similarly, he knows that waiters aiming for long-term profits are always fair ($ip_{Charlie,R1}(s_2) = \langle \textit{fair} \mapsto 1 \rangle$) and waiters aiming for short-term profit and reputation are sometimes fair, and sometimes greedy ($ip_{Charlie,R1}(s_3) = \langle \textit{fair} \mapsto 1/2, \textit{greedy} \mapsto 1/2 \rangle$). Charlie therefore suspects that Rob is aiming for short-term profit, but he cannot exclude the possibility of Rob aiming for short-term profits and reputation. His goal preference function on state s_0 is $gp_{Charlie,R1}(s_0) = \langle \textit{stp} \mapsto 4/5, \textit{stp+rep} \mapsto 1/5 \rangle$, where the values can be obtained by statistical inference from Charlie's past experience.

6) Beliefs: In partially observable systems, *belief* is often introduced to deal with agents' uncertainty about the current state of the system. Intuitively, belief expresses what agent *thinks* the current state of the execution is, and can be concretely represented as a probability distribution over states of the system. In ASMAS, beliefs of each agent are strongly related to their preference functions. For instance, assuming the preference functions defined above, after the first transition, Charlie's belief would be $\langle s_1 \mapsto 4/5, s_3 \mapsto 1/5 \rangle$, and after the second transition – $\langle s_5 \mapsto 4/5, s_8 \mapsto 1/10, s_9 \mapsto 1/10 \rangle$.

7) Iterated model variant: Finally, we briefly describe a possible extension of our scenario, which we call the Iterated Robotic Waiter Example. It involves introducing additional transitions to the existing model, so that the customer may order as many meals as they wish, as long as each new order comes after the previous order has arrived. Furthermore, following the delivery of an order, the waiter may change its goals and/or intentions. The extra transitions would therefore originate in the states which are terminal in the current model, and target each of states $s_4, \ldots, s_9$, corresponding to various cognitive states of the waiter. Extending the example in this way enables one to consider how trust between agents changes as a result of their interaction.

C. Expressing trust properties

To reason formally about properties of a given ASMAS, we use Probabilistic Rational Temporal Logic (PRTL*), which extends the well-known probabilistic temporal logic PCTL* [3] with cognitive and trust operators. Rather than giving full syntax and formal semantics of the language (we refer interested readers to [5]), we describe the intuitive meaning of selected operators and example formulas expressed in it. Throughout this section, ϕ denotes a state formula, whereas ψ denotes a path formula.

The first operator, $\mathbb{G}_A\phi$, expresses that ϕ holds in the future regardless of agent A changing its goals. For example:

- $\mathbb{G}_{waiter}\mathtt{P}^{\leq 0.9}\Diamond onTime$ – "Regardless of the waiter changing his goals, the probability of the meal arriving on time is no greater than 90%".

Similarly, $\bar{\mathbb{I}}_A\phi$ expresses that it is possible to achieve ϕ by changing agent A's intention. We may use it as follows:

- $\bar{\mathbb{I}}_{customer}\mathbb{G}_{waiter}\mathtt{P}^{\geq 0.5}\Diamond onTime$ – "The customer can change their intention, so that, regardless of the waiter changing his goals, the probability of the meal arriving on time is at least 50%".[1]

Next, $\mathbb{B}_A^{\bowtie q}\phi$, called the belief operator, states that agent A believes ϕ with probability in relation $\bowtie$ with q and can be used as follows:

- $\mathbb{B}^{\geq 0.5}_{customer}\Box(greedy_{waiter} \rightarrow \Box greedy_{waiter})$ – "The customer's belief that, once the waiter becomes greedy, he will remain greedy forever, is at least 50%".

Finally, we have two trust operators, $\mathbb{CT}^{\bowtie q}_{A,B}\psi$ and $\mathbb{DT}^{\bowtie q}_{A,B}\psi$. The first one, called *competence trust* operator, expresses that agent A trusts agent B with probability in relation $\bowtie$ with q on its capability of completing the task ψ. For example:

- $\mathbb{CT}^{\geq 0.7}_{customer,waiter}\Box(expMeal \rightarrow \mathtt{P}^{\geq 0.9}\Diamond onTime)$ – "The customer's trust in the waiter's capability of ensuring that the meal will arrive on time with probability at least 90% is greater or equal to 70%".

Second, $\mathbb{DT}^{\bowtie q}_{A,B}\psi$, referred to as the *disposition trust* operator, expresses that agent A trusts agent B with probability in relation $\bowtie$ with q on its willingness to do the task ψ We may combine belief and trust operators in the following way:

- $\mathbb{B}^{\geq 0.8}_{waiter}\mathbb{DT}^{\geq 0.7}_{customer,waiter} greedy_{waiter}$ – "The waiter's belief that the customer has at least 70% trust in him being greedy is at least 80%".

We note that the Robotic Waiter Example has been modelled using an extension of the PRISM model checker, which supports partially observable systems [7]. Even though the tool does not support PRTL* as the specification language, some properties may be verified by adapting the model and using PRISM's property specification language.

III. Conclusion

We have presented the overview of the concepts underlying Autonomous Stochastic Multi-Agent Systems and given an insight into Probabilistic Rational Temporal Logic by modelling the Robotic Waiter Example and considering its properties. The basic setting we considered in this paper allowed us to highlight the most important notions inherent in ASMAS.

Acknowledgments

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

[1]For simplicity, we have not defined customers' intentions in this paper, but our model could be easily extended to account for them.

REFERENCES

[1] Rino Falcone and Cristiano Castelfranchi. Social trust: A cognitive approach. In *Trust and deception in virtual societies*, pages 55–90. Springer, 2001.

[2] Gavin Haines. a taste of the future? meet singapores robotic waitresses. *The Telegraph*, Sep. 2016. URL http://www.telegraph.co.uk/travel/destinations/asia/singapore/articles/meet-singapores-robotic-waitresses. [Online; posted 7-September-2016; http://www.telegraph.co.uk/travel/destinations/asia/singapore/articles/meet-singapores-robotic-waitresses].

[3] Hans Hansson and Bengt Jonsson. A logic for reasoning about time and reliability. *Formal aspects of computing*, 6(5):512–535, 1994.

[4] Russell Hardin. *Trust and trustworthiness*. Russell Sage Foundation, 2002.

[5] Xiaowei Huang and Marta Kwiatkowska. Reasoning about cognitive trust in stochastic multiagent systems. In *Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, 2017.

[6] Dave Lee. US opens investigation into tesla after fatal crash. *British Broadcasting Corporation (BBC) News*, Jul. 2016. URL http://www.bbc.co.uk/news/technology-36680043. [Online; posted 1-July-2016; http://www.bbc.co.uk/news/technology-36680043].

[7] G. Norman, D. Parker, and Xueyi Zou. Verification and control of partially observable probabilistic systems. *Real-Time Systems*, 53(3):354–402, 2017.