Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Reasoning about Cognitive Trust in Stochastic Multiagent Systems*

Xiaowei Huang, Marta Kwiatkowska
Department of Computer Science
University of Oxford

Abstract

We consider the setting of stochastic multiagent systems and formulate an automated verification framework for quantifying and reasoning about agents' trust. To capture human trust, we work with a cognitive notion of trust defined as a subjective evaluation that agent A makes about agent B's ability to complete a task, which in turn may lead to a decision by A to rely on B. We propose a probabilistic rational temporal logic PRTL*, which extends the logic PCTL* with reasoning about mental attitudes (beliefs, goals and intentions), and includes novel operators that can express concepts of social trust such as competence, disposition and dependence. The logic can express, for example, that "agent A will eventually trust agent B with probability at least p that B will behave in a way that ensures the successful completion of a given task". We study the complexity of the automated verification problem and, while the general problem is undecidable, we identify restrictions on the logic and the system that result in decidable, or even tractable, subproblems.

Introduction

Mobile autonomous robots are rapidly entering the fabric of our society, to mention driverless cars and home assistive robots. Since robots are expected to work with or alongside humans in our society, they need to form partnerships with humans, as well as other robots, understand the social context, and behave, and be seen to behave, according to the norms of that context. Human partnerships such as cooperation are based on *trust*, which is influenced by a range of subjective factors that include preferences and experience. As the degree of autonomy of mobile robots increases and the nature of partnerships becomes more complex, and includes also shared autonomy, understanding and reasoning about social trust and the role it plays in decisions whether to rely on autonomous systems is of paramount importance. A pertinent example is the recent Tesla fatal car accident while on autopilot mode (Lee 2016), which is a result of over-reliance ("overtrust") by the driver, likely influenced through his personal motivation and preferences.

Trust is a complex notion, viewed as a belief, attitude, intention or behaviour, and is most generally understood as *a subjective evaluation of a truster on a trustee about something in particular*, e.g., the completion of a task (Hardin 2002). A classical definition from organisation theory (Mayer, Davis, and Schoorman 1995) defines trust as *the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party*. The importance of being able to correctly evaluate and calibrate trust to guide reliance on automation was recognised by Lee and See (2004). Trust (and trustworthiness) have also been actively studied in many application contexts such as security (Kagal, Finin, and Joshi 2001) and e-commerce (Corbitt, Thanasankit, and Yi 2003). However, in this paper we are interested in trust that governs social relationships between humans and autonomous systems, and therefore consider *cognitive trust* that captures the human notion of trust.

Human trust is influenced by culture and evolves over time. It is updated similarly to belief, based on social interactions and past experience, and increases in response to positive experiences but and may be negatively affected by failures. Trust is also based on opinions, which entails the need to incorporate subjective judgment that may differ for each agent depending on their individual experience. By understanding how human trust in an autonomous system evolves and being able to reason about it, we offer guidance for selecting an appropriate level of reliance on autonomy. We can also provide means to explain trust-based decisions.

We therefore aim to develop foundations for automated *quantitative* reasoning about trust between (human and robotic) agents, which can be employed to support decision making in dynamic, stochastic environments endowed with cognitive architecture. We work in the setting of stochastic multiagent systems, where stochasticity can be used to model, e.g., component failure or environmental uncertainty, and agents are endowed with individual goals and preferences. Inspired by the concepts of social trust of Falcone and Castelfranchi (2001), we formulate a probabilistic rational temporal logic PRTL* as a combination of the probabilistic temporal logic PCTL* (Hansson and Jonsson 1994; Bianco and de Alfaro 1995) with cognitive attitude operators (belief, goal, intention and capability) and a collection of novel trust operators (competence, disposition and dependence). The logic is able to express properties informally de-

*This work is supported by ERC AdG VERIWARE and EPSRC Mobile Autonomy Programme Grant EP/M019918/1.

fined by Falcone and Castelfranchi (2001) such as "agent A will eventually trust agent B with probability at least p that B will behave in a way that ensures the successful completion of a given task". The logic is interpreted over a stochastic multiagent system, where the cognitive reasoning processes for each agent can be modelled based on a *cognitive mechanism* that describes agent's mental state (a set of goals and an intention) and subjective preferences.

Since we wish to model dynamic evolution of beliefs and trust, the mechanisms are history-dependent, and thus the underlying semantics is an infinite branching structure, where we distinguish between agents' transitioning along the *temporal* and *cognitive* dimensions, as opposed to the classical accessibility relation employed in logics for agency. We resolve nondeterminism in the cognitive dimension by *preference functions*, given as probability distributions that model subjective knowledge about other agents. This is similar to quantification over adversaries (Halpern and Tuttle 1993) employed for systems exhibiting both nondeterminism and probability that results in fully probabilistic systems. Also, in contrast to defining beliefs in terms of knowledge and probabilistic knowledge (Halpern and Tuttle 1993) operators, which are based solely on agents' (partial) observations, we additionally allow agents' changes of mental state and subjective preferences to influence their belief.

Motivated by the need to evaluate trust in human-robot partnerships to support autonomous decision making, we aim to provide methods and software tools for reasoning about trust. We therefore study the foundations of the model checking problem for PRTL*. Unsurprisingly, we find that the general problem is undecidable. We then identify restrictions on the logic and the system that result in practically relevant yet decidable, and even tractable, subproblems. These decidable fragments can guide the design of a language to express cognitive trust and enable practical implementation and validation of the techniques.

This paper presents a simplified framework for a single agent that does not support nested beliefs and trust. An accompanying technical report (Huang and Kwiatkowska 2016) contains a comprehensive treatment of the framework, including the full logic, a connection with strategic reasoning, illustrative examples and full proofs.

Related Work

The notion of trust has been widely studied in management, psychology, philosophy and economics (see (Lahijanian and Kwiatkowska 2016) for an overview). Trust in the context of human-technology relationships can be roughly classified into three categories: *credentials-based*, *experience-based*, and *cognitive trust*. Credentials-based trust is used mainly in security, where a user must supply credentials in order to gain access. Experience-based trust, which includes reputation-based trust in peer-to-peer and e-commerce applications, involves online evaluation of a trust value for an agent informed by experiences of interaction with that agent. A formal foundation for quantitative reputation-based trust has been proposed by Krukow, Nielsen, and Sassone (2008). In contrast, we focus on (quantitative) cognitive trust, which captures the social (human) notion of trust and, in particular, trust-based decisions between humans and robots.

The cognitive trust distinguishes itself from the other trust notions by explicitly accounting for not only the human experience, but also subjective judgement about preferences and the mental state of other agents. The cognitive theory of Falcone and Castelfranchi (2001), itself founded on organisational trust of Mayer, Davis, and Schoorman (1995), provides an intuitive definition of complex trust notions but lacks rigorous semantics. Several papers, e.g., (Jøsang 2001; Herzig et al. 2010; Herzig, Lorini, and Moisan 2013), have formalised the theory of Falcone and Castelfranchi (2001) using modal logic, but none are quantitative and automatic verification is not considered. Of relevance are recent approaches (Sweet et al. 2016; Setter, Gasparri, and Egerstedt 2016) that model the evolution of trust in human-robot interactions as a dynamical system; instead, our formalism supports evolution of trust through events and agent interactions.

A number of logic frameworks have been proposed that develop the theory of human decisions (Bratman 1987) for artificial agents, see (Meyer, Broersen, and Herzig 2014) for a recent overview. The main focus has been on studying the relationships between modalities with various axiomatic systems, but their amenability to automatic verification is arguable because of a complex underlying possible world semantics, to mention the sub-tree relation of BDI logic. The only attempt at model checking such logics (Schild 2000) ignores the internal structure of the possible worlds to enable a reduction to temporal logic model checking.

The distinctive aspect of our work is thus a *quantitative* formalisation of *cognitive* trust in terms of *probabilistic* temporal logic, based on a Bayesian notion of *belief*, together with algorithmic complexity of the corresponding *model checking* problem.

Cognitive Theory of Social Trust

In the context of automation, trust is understood as delegation of responsibility for actions to the autonomous system and willingness to accept risk (possible harm) and uncertainty. The decision to delegate is based on a *subjective* evaluation of the system's capabilities for a particular task, informed by factors such as past experience, social norms and individual preferences. Moreover, trust is a dynamic concept which evolves over time, influenced by events and past experience. The cognitive processes underpinning trust are captured in the influential theory of social trust by Falcone and Castelfranchi (2001), which is particularly appropriate for human-robot relationships and serves as an inspiration for this work.

Falcone and Castelfranchi (2001) 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. They consider agent A's trust in agent B for a specific goal ψ (goals may be divided into tasks), and distinguish the following core concepts: *competence* trust, where A believes that B is able to perform ψ, and *disposition* trust, where A believes that B is willing to perform ψ. The decision to delegate or rely on B involves a

complex notion of trust called *dependence*: A believes that A needs, depends, or is at least better off to rely on B to achieve ψ, which has two forms, *strong* (A needs or depends on B) and *weak* (for A, it is better to rely than not to rely on B). Falcone and Castelfranchi (2001) also identify *fulfillment* belief arising in the truster's mental state, which we do not consider for simplicity.

We therefore work in a stochastic setting (to represent uncertainty), aiming to *quantify* belief probabilistically and express trust as a *subjective, belief-weighted expectation*, which is informally understood as a degree of trust.

Autonomous Stochastic Multiagent Systems

We work with the notion of stochastic multiagent systems, which can express beliefs and goals of individual agents, as well as random events to capture uncertainty. Before we define the cognitive aspects, we introduce stochastic games as the underlying behavioural model. Let $\mathcal{D}(X)$ denote the set of probability distributions on a set X.

Definition 1 *A* stochastic multiplayer game (SMG) *is a tuple* $M = (Ags, S, s_{\text{init}}, \{Act_A\}_{A\in Ags}, T)$*, where* Ags *is a finite set of agents,* S *is a finite set of states,* $s_{\text{init}} \in S$ *is an initial state,* Act_A *is a finite set of actions for agent* A*, and* $T : S \times Act \to \mathcal{D}(S)$ *is a (partial) transition probability function such that* $Act = \times_{A\in Ags} Act_A$.

We refer to players of a game as agents. Let a_A be agent A's action in the joint action $a \in Act$. We let $Act(s) = \{a \in Act \mid T(s,a) \text{ is defined}\}$ and $Act_A(s) = \{a_A \mid a \in Act(s)\}$. For technical reasons, we assume that $Act(s) \neq \emptyset$ for all $s \in S$; this is not a restriction, as any system can be transformed to one that satisfies this condition.

States S are global, and encode agents' local states as well as environment states. In each state s, agents independently (and possibly at random) choose a local action (which may include the silent action $\perp$), the environment performs an update, and the system transitions to a state s' satisfying $T(s,a)(s') > 0$, where a is the *joint* action. Sometimes we also denote transitions by $s \longrightarrow^{a}_{T} s'$ with the meaning $\exists s'.T(s,a)(s') > 0$.

We define a finite, resp. infinite, *path* ρ as a sequence of states $s_0 s_1 s_2 ...$ such that $T(s_i, -)(s_{i+1}) > 0$ for all $i \geq 0$, and denote the set of finite and infinite paths of M starting in s, respectively, by $\text{FPath}^M_T(s)$ and $\text{IPath}^M_T(s)$, and sets of paths starting from any state by FPath^M_T and IPath^M_T, and omit M if clear from context. For a finite path ρ we write $\text{last}(\rho)$ to denote the last state. We refer to paths induced from the transition probability function T as the *temporal dimension*.

For an agent A we define an *action strategy* σ_A as a function $\sigma_A : \text{FPath}^M_T \longrightarrow \mathcal{D}(Act_A)$ such that for all $a_A \in Act_A$ and finite path ρ it holds that $\sigma_A(\rho)(a_A) > 0$ only if $a_A \in Act_A(\text{last}(\rho))$. An *action strategy profile* σ is a vector of action strategies $(\sigma_A)_{A\in Ags}$. Under a fixed σ, one can define a probability measure $\Pr^{M,\sigma}$ on $\text{IPath}^M_T(s_{\text{init}})$ in the standard way.

Beliefs

Since full knowledge of states of other agents is not a realistic assumption, we introduce partial observations.

Definition 2 *A* partially observable *stochastic multiplayer game (poSMG) is a tuple* $M = (Ags, S, s_{\text{init}}, \{Act_A\}_{A\in Ags}, T, \{O_A\}_{A\in Ags}, \{obs_A\}_{A\in Ags})$*, where* $(Ags, S, s_{\text{init}}, \{Act_A\}_{A\in Ags}, T)$ *is an SMG,* O_A *is a finite set of observations for agent* A*, and* $obs_A : S \longrightarrow O_A$ *is a labelling of states with observations for agent* A *such that, for all* $A \in Ags$ *and* $s, s' \in S$*, we have* $obs_A(s) = obs_A(s')$ *implies* $Act_A(s) = Act_A(s')$.

Note that each agent is able to remember its past observations, known as synchronous perfect recall. Also note that we work with deterministic observations, but emphasise that stochastic observations, which are more realistic in presence of environmental uncertainty, can be added without increasing the complexity of the problem (Huang and Kwiatkowska 2016).

The notions of $\text{FPath}^M_T(s)$, $\text{IPath}^M_T(s)$, FPath^M_T and IPath^M_T generalise to poSMG. However, since states cannot be directly observed, action strategies are now based on observations, that is, for any finite paths $\rho = s_0 s_1 ... s_m$ and $\rho' = s'_0 s'_1 ... s'_m$ satisfying $obs_A(s_k) = obs_A(s'_k)$ for all $0 \leq k \leq m$ we have $\sigma_A(\rho) = \sigma_A(\rho')$. As before, a strategy profile σ induces a probability measure $\Pr^{M,\sigma}$ on $\text{IPath}^M_T(s_{\text{init}})$.

We employ a well known construction (Chatterjee, Chmelik, and Tracol 2016) which, for a poSMG M and a fixed agent A, induces a (possibly infinite state) *belief SMG* $Bel_A(M)$, whose states are *distributions* over states S of M called *belief states*. This construction can be effectively used to reason about one agent's understanding of the system, but does not support nested reasoning about agents' beliefs. For the general construction see (Huang and Kwiatkowska 2016).

Formally, a belief SMG for poSMG M and agent A is a tuple $Bel_A(M) = (Ags, \mathcal{D}(S), \beta_{init}, \{Act_A\}_{A\in Ags}, T^{Bel}_A)$, where β_{init} is an initial Dirac belief distribution over s_{init} and, for any $\beta, \beta' \in \mathcal{D}(S)$,

$$T^{Bel}_A(\beta, a)(\beta') = \sum_{s\in S} \beta(s) \cdot \Big(\sum_{o\in O_A \,\&\, \beta^{a,o}=\beta'} \; \sum_{s'\in S \,\&\, \beta'(s')>0} T(s,a)(s') \Big) \quad (1)$$

and $\beta^{a,o}$ is belief reached from β by performing a and observing o, i.e.,

$$\beta^{a,o}(s') = \begin{cases} \frac{\sum_{s\in S} T(s,a)(s')\cdot\beta(s)}{\sum_{s\in S}(\sum_{s''\in S \,\&\, obs_A(s'')=o} T(s,a)(s'')\cdot\beta(s))} & \text{if } obs_A(s') = o \\ 0 & \text{otherwise.} \end{cases}$$

We define a mapping δ_A from paths of M to paths of $Bel_A(M)$ as follows: $\delta_A(s_{\text{init}}) = \beta_{init}$, and $\delta_A(\rho s) = \delta_A(\rho)\beta'$ such that $\beta'(s) > 0$ and $T^{Bel}_A(\text{last}(\delta_A(\rho)), a)(\beta') > 0$ for some a. We note that T^{Bel}_A satisfies the constraints of a transition probability function and can be viewed as a Bayesian update. As before, a strategy profile σ induces a probability measure $\Pr^{Bel_A(M),\sigma}$ over infinite paths of SMG $Bel_A(M)$.

Cognitive mechanism

Since we wish to model mental attitudes in order to reason about trust, we endow agents with a *cognitive mechanism* inspired by the BDI framework (beliefs, desires and intentions) in the sense of (Bratman 1987). In addition to beliefs that we have just defined, we introduce the concepts of *goals* and *intentions* (also called pro-attitudes), as well as *subjective preferences*. The existence of (own) goals and intentions

is a key distinction made about autonomy. For an agent A, the idea is that, while actions Act_A represent A's actions in the *physical* space, pro-attitudes represent the *cognitive* processes that lead to decisions about which action to take. We thus distinguish two *dimensions* of transitions, temporal (behavioural) and cognitive.

Definition 3 *An* autonomous stochastic multiagent system *(ASMAS) is a tuple* $M = (Ags, S, s_{\text{init}}, \{Act_A\}_{A\in Ags}, T, \{O_A\}_{A\in Ags}, \{obs_A\}_{A\in Ags}, \{\Omega_A\}_{A\in Ags})$, *where* $(Ags, S, s_{\text{init}}, \{Act_A\}_{A\in Ags}, T, \{O_A\}_{A\in Ags}, \{obs_A\}_{A\in Ags})$ *is a poSMG and* $\Omega_A = (\{Goal_A\}_{A\in Ags}, \{Int_A\}_{A\in Ags}, \{gp_{A,B}\}_{A,B\in Ags}, \{ip_{A,B}\}_{A,B\in Ags})$ *is a cognitive mechanism, where* $Goal_A$ *is a finite set of goals for agent A;* Int_A *is a finite set of intentions for agent A;* $gp_{A,B} : S \longrightarrow \mathcal{D}(2^{Goal_B})$ *assigns to each state, from A's point of view, a distribution over possible goal changes of B; and* $ip_{A,B} : S \longrightarrow \mathcal{D}(Int_B)$ *assigns to each state, from A's point of view, a distribution over possible intentional changes of B.*

An agent can have several goals, not necessarily consistent, but only a single intention. Goals and intentions are abstract entities. We think of goals as abstract attitudes, for example selflessness or risk-taking, whereas intentions are concretely implemented in our (simplified) setting as *action strategies*, thus identifying the next (possibly random) action to be taken in the temporal dimension.

We extend the set of temporal transitions $s \longrightarrow_T^a s'$ with cognitive transitions for agent A corresponding to a change of goal (respectively intention) to x, denoted $s \longrightarrow_C^{A.g.x} s'$ if x is the goal set for A in s' (respectively $s \longrightarrow_C^{A.i.x} s'$ if x is the intention of s'). It is noted that $s \longrightarrow_C^{B.g.x} s'$ for some s' only when $gp_{A,B}(s)(x) > 0$, and $s \longrightarrow_C^{B.i.x} s'$ for some s' only when $ip_{A,B}(s)(x) > 0$. We extend the transition probability function T in the obvious way by letting, e.g., $T(s, A.g.x)(s') = 1$ when $s \longrightarrow_C^{A.g.x} s'$, and $T(s, A.g.x)(s') = 0$ otherwise. We denote by $\text{FPath}^M(s)$, $\text{IPath}^M(s)$, FPath^M and IPath^M the sets of paths formed by extending the sets $\text{FPath}_T^M(s)$, $\text{IPath}_T^M(s)$, FPath_T^M and IPath_T^M of temporal paths with paths thet interleave the cognitive and temporal transitions.

To simplify the presentation, we make a *deterministic behaviour assumption* on ASMAS by requiring that for each state s there exists a unique joint action $a \in Act$ such that a (non-unique) state s' is chosen with probability $T(s, a)(s')$. The intuition is that mental states determine temporal actions. We emphasise, though, that our framework can be generalised (Huang and Kwiatkowska 2016) to allow also strategic reasoning (quantification over strategies) in the sense of in the style of (Chen et al. 2013).

The deterministic behaviour assumption resolves the nondeterminism in the temporal dimension and removes the need to quantify over action strategies. However, potential changes to pro-attitudes introduce another source of nondeterminism, which must be resolved in order to define probability measure over infinite paths $\text{IPath}^M(s_{\text{init}})$. We achieve this by defining *cognitive reasoning strategies* g_A and i_A, which are *history dependent* and model *subjective* preferences of A. Formally, we define the cognitive goal strategy as $g_A : \text{FPath} \longrightarrow \mathcal{D}(2^{Goal_A})$, and the intentional strategy as $i_A : \text{FPath} \longrightarrow \mathcal{D}(Int_A)$. While we do not discuss how one arrives at such a strategy, we remark that they may result from reasoning supported by cognitive architectures, with the subjective view induced by goal and intentional *preference functions*, $gp_{A,B}$ and $ip_{A,B}$. The key idea behind the preference functions is that they replace preference ordering of BDI reasoning, and instead model *probabilistic prior* knowledge of agent A about goals and intentions of B, informed by prior experience (through observations) and aspects such as personal preferences and social norms, which may vary for different cultures. We also need to assume that goal and intentional strategies are based on observations, that is, for any finite paths $\rho = s_0 s_1 ... s_m$ and $\rho' = s'_0 s'_1 ... s'_m$ satisfying $obs_A(s_k) = obs_A(s'_k)$ for all $0 \le k \le m$ we have $g_A(\rho) = g_A(\rho')$, and similarly for i_A.

Formally, given an ASMAS M and an agent A, we work in the induced belief SMG $Bel_A(M)$, where the transition probability function T_A^{Bel} is redefined to take into account cognitive transitions as follows. Note that the mapping δ_A implies that, for each finite path ρ, there exists a unique β such that $\beta = \text{last}(\delta_A(\rho))$, and we write β_ρ for such a belief state. For any $\beta' \in \mathcal{D}(S)$, we define

$$T_A^{Bel}(\beta_\rho, A.g.x)(\beta') = \sum_{s\in S} \beta_\rho(s) \cdot \Big(\sum_{o\in O_A \,\&\, (\beta_\rho)^{A.g.x.o} = \beta'} \sum_{s'\in S \,\&\, \beta'(s')>0} g_A(\rho)(x) \cdot T(s, A.g.x)(s')\Big) \tag{2}$$

and $(\beta_\rho)^{A.g.x.o}$ is the belief state reached from β_ρ by A performing goal change into x and observing o, i.e., $(\beta_\rho)^{A.g.x.o}(s') =$

$$\begin{cases} \frac{\sum_{s\in S} g_A(\rho)(x)\cdot\beta_\rho(s)\cdot T(s,A.g.x)(s')}{\sum_{s\in S}(\sum_{x'} g_A(\rho)(x')\cdot\beta_\rho(s)\cdot T(s,A.g.x')(s'))} & \text{if } obs_A(s') = o \\ 0 & \text{otherwise.} \end{cases}$$

The construction is similar for intentional changes $T_A^{Bel}(\beta_\rho, A.i.x)$. Cognitive transitions of agents other than A can be handled in a similar way by using functions $gp_{A,B}$ and $ip_{A,B}$ instead of g_A and i_A in the above expressions. Then, under the pair of strategies g_A and i_A, the modified transition probability function T_A^{Bel} induces a probability measure $\Pr^{Bel_A(M), g_A, i_A, \Omega_A}$ over infinite paths of SMG $Bel_A(M)$. Note that the construction of T_A^{Bel} considers agent A's cognitive reasoning strategies and preference functions, but not other agents' cognitive strategies: this is to avoid cyclic reasoning which can be complicated both conceptually and computationally. We emphasise that our approach is not limited to a single agent, see (Huang and Kwiatkowska 2016).

Example 1 *We consider a simple trust game from (Kuipers 2016) involving two agents, Alice and Bob. At the beginning, Alice has \$10 and Bob has \$5. If Alice does nothing, then everyone keeps what they have. If Alice invests her money with Bob, then Bob can turn the \$15 into \$40 and then decide whether to share the investment yield with Alice. If so, each will have \$20 and, otherwise, Alice will lose her money and Bob gets \$40. The game has Nash equilibria of Alice withholding her money and Bob keeping the yield. Under the standard economic assumptions of rational self-interest, the predicted behaviour for Alice is to withhold the money. This behaviour is not reproduced in experiments, with most human players willing to invest, which can be explained by explicitly considering trust in economic decisions.*

Probabilistic Rational Temporal Logic

We introduce a logic PRTL* to express properties of agents in autonomous stochastic multiagent systems. PRTL* combines the probabilistic temporal logic PCTL* with operators for reasoning about agents' beliefs and cognitive trust. As suggested by Falcone and Castelfranchi (2001), we express trust in terms of belief, which probabilistically quantifies the degree of trust as a function of subjective certainty, e.g., "I am 99% certain that the autonomous taxi service is trustworthy", or "I trust the autonomous taxi service 99%". The logic captures how the value of 99% can be computed based on the agent's past experience and (social, economic) preferences.

Definition 4 *The syntax of the language PRTL* is:*

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

where p is an atomic proposition, $A, B \in Ags$, $\bowtie \in \{<, \leq, >, \geq\}$, and $q \in [0, 1]$.

In the above, ϕ is a PRTL* formula and ψ an LTL (path) formula. The operator $\forall$ is the path quantifier of CTL* and $\mathrm{P}^{\bowtie d}\psi$ is the probabilistic operator of PCTL (Hansson and Jonsson 1994), which denotes the probability of those future infinite paths that satisfy ψ, evaluated in the temporal dimension. We omit the description of standard and derived ($\phi_1 \wedge \phi_2$, $\Diamond\psi$ and $\exists\phi$) operators, and just focus on the added operators. Although not explicitly defined, we can reason about rewards by assigning values to state variables and reasoning about those values. The model can be extended with rewards and the reward operator in the style of (Chen et al. 2013) can be added to the logic.

The *cognitive* operators $\mathbb{G}_A\psi$, $\mathbb{I}_A\psi$ and $\mathbb{C}_A\psi$ consider the task expressed as ψ and respectively quantify, in the cognitive dimension, over possible changes of goals, possible intentions and available intentions. Thus, $\mathbb{G}_A\psi$ expresses that ψ holds in future regardless of agent A changing its goals. Similarly, $\mathbb{I}_A\psi$ states that ψ holds regardless of A changing its (not necessarily available) intention, whereas $\mathbb{C}_A\psi$ quantifies over the available intentions, and thus expresses that agent A can change its intention to achieve ψ.

$\mathbb{B}_A^{\bowtie q}\psi$ is the *belief* operator, which states that agent A believes ψ with probability in relation $\bowtie$ with q. In contrast to BDI logics, we work with quantitative beliefs. We introduce operators for the two core trust concepts of Falcone and Castelfranchi (2001). $\mathbb{CT}_{A,B}^{\bowtie q}\psi$ is the *competence trust* operator, meaning that agent A trusts agent B with probability in relation $\bowtie$ with q on its capability of completing the task ψ, where capability is understood to be the existence of a valid intention (in $Int_B(s)$ for s being the current state) to implement the task. $\mathbb{DT}_{A,B}^{\bowtie d}\psi$ is the *disposition trust* operator, which expresses that agent A trusts agent B with probability in relation $\bowtie$ with q on its willingness to do the task ψ, where the state of willingness is interpreted as that the task is unavoidable for all intentions in intentional strategy (i.e., $i_B(\rho)$ for ρ being the path up to the current point in time).

Example 2 *For the trust game example, the formula*

$$\mathbb{DT}_{Alice,Bob}^{\geq 0.7} sharing_{Bob}$$

where $sharing_{Bob}$ is an atomic proposition, expresses that Alice's trust in Bob's willingness to share the profit is at least 70%, and

$$\mathbb{B}_{Bob}^{\geq 0.8}\mathbb{DT}_{Alice,Bob}^{\geq 0.7} investor_{Bob}$$

states that Bob's belief that Alice has at least 70% trust in him being an investor is at least 80%, where $investor_{Bob}$ is an atomic proposition.

Let $\beta \in \mathcal{D}(S)$ be a belief state. For a measurable function $f : S \longrightarrow [0, 1]$, we denote by $E_\beta[f]$ the belief-weighted expectation of f, i.e., $E_\beta[f] = \sum_{s\in S} \beta(s) \cdot f(s)$. We also let

$$Prob(M, \rho, \psi) \equiv \Pr\{\delta \in \mathrm{IPath}_T^M(\mathrm{last}(\rho)) \mid M, \rho, \delta \models \psi\}$$

to denote the probability of satisfying ψ in future (note that the future only concerns the temporal dimension). We also write $B.i(s, x)$ to denote the (unique) state s' such that $s \longrightarrow_C^{B.i.x} s'$, and similarly for $B.g(s, x)$.

We now define semantics for PRTL* for the simplified setting of formulas without nested belief and trust, where we omit the cases where standard definitions apply for reasons of space. The reader is referred to (Huang and Kwiatkowska 2016) for the complete semantics.

Definition 5 *Let (M, AP, L) be a labelled ASMAS where M is an ASMAS satisfying a deterministic behaviour assumption with fixed preference functions and $L : S \longrightarrow 2^{AP}$ is a labelling with atomic propositions. $supp(f)$ denotes the support of the distribution f. The semantics of the logic PRTL* is defined by a relation $M, \rho \models \phi$ for $\rho \in \mathrm{FPath}^M$, a finite path of M, inductively over the structure of the formula ϕ as follows.*

- $M, \rho \models \mathrm{P}^{\bowtie q}\psi$ *if* $Prob(M, \rho, \psi) \bowtie q$
- $M, \rho s \models \mathbb{G}_A\psi$ *if* $\forall x \in supp(g_A(\rho s))\, \exists s' \in S : s \longrightarrow_C^{A.g.x} s'$ *and* $M, \rho s s' \models \psi$
- $M, \rho s \models \mathbb{I}_A\psi$ *if* $\forall x \in supp(i_A(\rho s))\, \exists s' \in S : s \longrightarrow_C^{A.i.x} s'$ *and* $M, \rho s s' \models \psi$
- $M, \rho s \models \mathbb{C}_A\psi$ *if* $\exists x \in Int_A(s)\, \exists s' \in S : s \longrightarrow_C^{A.i.x} s'$ *and* $M, \rho s s' \models \psi$
- $M, \rho \models \mathbb{B}_A^{\bowtie q}\psi$ *if*

$$E_{\mathrm{last}(\delta_A(\rho))}[\mathrm{Sat}_\psi] \bowtie q$$

where $\mathrm{Sat}_\psi(s) = 1$ *if* $M, s \models \psi$ *and* 0 *otherwise. Note that* $E_{\mathrm{last}(\delta_A(\rho))}[\mathrm{Sat}_\psi]$ *is defined because* $\mathrm{last}(\delta_A(\rho))$ *is a belief state for which expectation is defined.*

- $M, \rho \models \mathbb{CT}_{A,B}^{\bowtie q}\psi$ *for* $\bowtie \in \{\geq, >\}$ *if*

$$E_{\mathrm{last}(\delta_A(\rho))}[V_{\mathbb{CT},B,\psi}] \bowtie q$$

where

$$V_{\mathbb{CT},B,\psi}(s) = \sup_{x \in Int_B(s)} Prob(M, B.i(s, x), \psi)$$

and if $\bowtie$ *is* $\leq$ *or* $<$ *we replace* sup *with* inf *in the above.*

- $M, \rho \models \mathbb{DT}_{A,B}^{\bowtie d}\psi$ *if for* $\bowtie \in \{\geq, >\}$ *if*

$$E_{\mathrm{last}(\delta_A(\rho))}[V_{\mathbb{DT},B,\psi}] \bowtie q$$

where

$$V_{\mathbb{DT},B,\psi}(s) = \inf_{x \in i_B(\rho') \& \delta_A(\rho') = \delta_A(\rho) \& \mathrm{last}(\rho') = s} Prob(M, B.i(s, x), \psi)$$

and if $\bowtie$ *is* $\leq$ *or* $<$ *we replace* inf *with* sup *in the above.*

We interpret formulas ϕ in ASMAS M in a state reached after executing a path ρ, in history-dependent fashion. Note that this path ρ may have interleaved cognitive and temporal transitions, and has a corresponding path $\delta_A(\rho)$ in the belief ASMAS $Bel_A(M)$ that ends in the belief state $\text{last}(\delta_A(\rho))$. The cognitive operators quantify over possible changes of goals and intentions in M in the cognitive dimension only, reflecting the cognitive reasoning processes leading to a decision. The probabilistic operator computes the probability of future paths satisfying ψ (i.e. completing the task ψ) in M in the temporal dimension as for PCTL*, reflecting the physical actions resulting from the cognitive decision, and compares this to the probability bound q. The belief operator is evaluated in the belief ASMAS, and corresponds to the probability of satisfying ψ in future in the original ASMAS M weighted by the belief reached by following $\delta_A(\rho)$; in other words, it is a belief-weighted expectation of future satisfaction of ψ, which is subjective, as it is influenced by A's prior knowledge about B encoded in the preference function. The competence trust operator reduces to the computation of optimal probability of satisfying ψ in M over possible changes of agent's intention, which is again weighted by the belief $\text{last}(\delta_A(\rho))$ and compared to the probability bound q. Dispositional trust, on the other hand, computes the optimal probability of satisfying ψ in M over possible states of agent's willingness, which is weighted by the belief $\text{last}(\delta_A(\rho))$ and compared to the probability bound q.

The operators $\mathbb{CT}$ and $\mathbb{DT}$ cannot be derived in the logic PRTL* but reduce to $\mathbb{B}$ when ASMAS M has sure beliefs. Strong and weak dependence trust notions of Falcone and Castelfranchi (2001), though, can be modelled.

Definition 6 *We can introduce operators $\mathbb{ST}^{\bowtie q}_{A,B}$ and $\mathbb{WT}^{\bowtie}_{A,B}$ to express strong and weak dependence, respectively, with the semantics:*

- $M, \rho \models \mathbb{ST}^{\bowtie q}_{A,B}\psi$ *if* $M, \rho \models \mathbb{DT}^{\bowtie q}_{A,B}\psi \wedge \neg \mathbb{B}^{\bowtie q}_{A}\psi$
- $M, \rho \models \mathbb{WT}^{\bowtie}_{A,B}\psi$ *if*

$$E_{\text{last}(\delta_A(\rho))}[V_{\mathbb{DT},B,\psi}] \bowtie E_{\text{last}(\delta_A(\rho))}[V_{M,\psi}]$$

where $V_{M,\psi}(s) = Prob(M, s, \psi)$.

Strong dependence means that A depends on B to achieve ψ (i.e. ψ can be implemented through intentional change of B), which cannot be achieved otherwise (expressed as a belief in impossibility of ψ in future), and weak dependence that A is better off relying on B compared to doing nothing (meaning intentional changes of B can bring about better outcomes).

It is known that Nash equilibria may lead to undesirable solutions which discourage collaborations. Our framework can complement this by naturally allowing agents to take into consideration each other's trust when taking decisions. The agent can thus evaluate another agent's trust in him/her and then take the action to maintain or increase the other agent's trust to pursue long term collaboration, or take the action only when it believes that the other agent' trust meets a specific threshold. We illustrate this with the trust game example below.

Example 3 *Consider again the trust game example. In (Kuipers 2016), the computation of the payoff is amended to also include an estimate of trust that an agent has, namely the payoff is +5 when Bob is sharing and -20 when he keeps the yield. The new Nash equilibrium is for Alice to invest her money and Bob to share the investment yield. We can achieve the same result by modelling the game as an ASMAS, see (Huang and Kwiatkowska 2016) for the details. More specifically, we model the evolution of trust based on interactions and prior knowledge, whereby Alice's trust in Bob increases if he shares the yield, and decreases if he keeps it without sharing. Alice guards her decision whether to invest by considering if she has sufficient level of trust in Bob's willingness to share, e.g. $\mathbb{DT}^{\geq 0.7}_{Alice,Bob} sharing_{Bob}$. Thus, if Alice has prior positive experience of Bob's willingness to share she will be inclined to invest, and hence our notion of social trust helps to explain cases where actual human behaviour is at variance with standard economic and rationality theories.*

The precise value of the threshold for trust is context-dependent. The trust value higher than an appropriately calibrated level is known as 'overtrust'.

Example 4 *Based on his investment track record, Bob has been ranked by an agency to have the trust value of q. Then by model checking the property $\mathbb{DT}^{>q}_{Alice,Bob} investor_{Bob}$ we can determine whether Alice is overtrusting Bob. On the other hand, the property $\mathrm{P}^{\geq 1}\Diamond\mathbb{DT}^{\geq q}_{Alice,Bob} investor_{Bob}$ states that almost surely Alice will eventually trust Bob with probability at least q.*

We remark that our framework is more general than trust games, and can additionally capture personal goals and preferences, and how dynamic changes in these characteristics influence trust-based decisions.

Model Checking Complexity

Our automated verification framework accepts as inputs an ASMAS M and a PRTL* specification formula ϕ, and determines whether $M, s_{\text{init}} \models \phi$. Unfortunately, the problem is undecidable in general.

Theorem 1 *Model checking PRTL* is undecidable, even for formulas concerning beliefs of a single agent.*

In the following, we present three decidable fragments of the general problem. The first fragment, named BPRTL*, can express events occurring in a bounded number of steps. It allows formulas which (1) do not contain temporal operators $\mathtt{U}$ and $\Box$, (2) all X operators are immediately prefixed with a probabilistic operator or a branching operator, i.e., in a combination of $\mathrm{P}^{\bowtie q}\bigcirc\psi$, $\forall\bigcirc\psi$, or $\exists\bigcirc\psi$, and (3) the nested depth of belief and trust operators is constant.

Theorem 2 *The complexity of verifying the fragment BPRTL* is in EXPTIME and PSPACE-hard.*

The second fragment, named PRTL$^*_1$, allows the operators $\mathtt{U}$ and $\Box$, and has the full expressiveness of LTL, but is subject to restrictions on the model and formulas, and specifically (1) works with a single agent's beliefs and trust, (2) there are no nested beliefs or trust formulas, (3) beliefs and trust cannot be in the scope of a probabilistic operator P, and (4) there is a constant number of belief or trust operators.

Theorem 3 *The complexity of verifying the fragment $PRTL^*_1$ is PSPACE-complete.*

We also identify a fragment in which the belief or trust operators can be in the scope of a probabilistic operator but need to be qualitative, i.e., almost sure satisfaction. This fragment, named $PQRTL^*_1$, is, very surprisingly, polynomial time. Specifically, we restrict formulas to be of the form $\Box(\psi \Rightarrow \mathrm{P}^{\bowtie q}\Diamond\mathbb{B}^{\geq 1}_{A}\psi)$, $\Box(\psi \Rightarrow \mathrm{P}^{\bowtie q}\Diamond\mathbb{CT}^{\geq 1}_{A,B}\psi)$ or $\Box(\psi \Rightarrow \mathrm{P}^{\bowtie q}\Diamond\mathbb{DT}^{\geq 1}_{A,B}\psi)$ such that, in ψ, there are no belief or trust operators and every temporal operator is immediately prefixed with a branching operator, i.e., in the style of CTL. The system M needs to satisfy the formula $M \models \Box(\psi \Rightarrow \Box\psi)$, which means that, once ψ holds, it will hold henceforth.

Theorem 4 *The complexity of verifying the fragment $PQRTL^*_1$ is in PTIME.*

Conclusion

The paper proposes an automated verification framework for autonomous stochastic multiagent systems and specifications given in probabilistic rational temporal logic PRTL*, which includes novel modalities for quantifying and reasoning about agents' cognitive trust. We study computational complexity of the decision problems and show that, although the general problem is undecidable, there are decidable, even tractable, fragments. While preliminary notions of cognitive trust were proposed by Falcone and Castelfranchi (2001), the paper provides the first rigorous formalisation and the corresponding model checking procedure. Future work will include the definition of a Bellman operator to evaluate trust, integration with cognitive reasoning frameworks, and a tool implementation of the techniques.

References

Bianco, A., and de Alfaro, L. 1995. Model checking of probabalistic and nondeterministic systems. In *FSTTCS 1995*, 499–513.

Bratman, M. 1987. *Intentions, Plans, and Practical Reason*. Harvard University Press, Massachusetts.

Chatterjee, K.; Chmelik, M.; and Tracol, M. 2016. What is decidable about partially observable Markov decision processes with ω-regular objectives. *J. Comput. Syst. Sci.* 82(5):878–911.

Chen, T.; Forejt, V.; Kwiatkowska, M. Z.; Parker, D.; and Simaitis, A. 2013. Automatic verification of competitive stochastic systems. *Formal Methods in System Design* 43(1):61–92.

Corbitt, B. J.; Thanasankit, T.; and Yi, H. 2003. Trust and e-commerce: a study of consumer perceptions. *Electronic Commerce Research and Applications* 2(3):203–215.

Falcone, R., and Castelfranchi, C. 2001. Social trust: A cognitive approach. In *Trust and Deception in Virtual Societies*. Kluwer. 55–90.

Halpern, J. Y., and Tuttle, M. R. 1993. Knowledge, probability, and adversaries. *Journal of the ACM* 40(3):917–962.

Hansson, H., and Jonsson, B. 1994. A logic for reasoning about time and reliability. *Formal aspects of computing* 6(5):512–535.

Hardin, R. 2002. *Trust and trustworthiness*. Russell Sage Foundation.

Herzig, A.; Lorini, E.; Hübner, J. F.; and Vercouter, L. 2010. A logic of trust and reputation. *Logic Journal of the IGPL* 18(1):214–244.

Herzig, A.; Lorini, E.; and Moisan, F. 2013. A simple logic of trust based on propositional assignments. *The Goals of Cognition. Essays in Honor of Cristiano Castelfranchi*.

Huang, X., and Kwiatkowska, M. 2016. Reasoning about cognitive trust in stochastic multiagent systems. Technical Report CS-RR-16-02, University of Oxford.

Jøsang, A. 2001. A logic for uncertain probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 9(03):279–311.

Kagal, L.; Finin, T.; and Joshi, A. 2001. Trust-based security in pervasive computing environments. *Computer* 34(12):154 – 157.

Krukow, K.; Nielsen, M.; and Sassone, V. 2008. Trust models in ubiquitous computing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 366(1881):3781–3793.

Kuipers, B. 2016. What is trust and how can my robot get some? (presentation). In *RSS 2016 Workshop on Social Trust in Autonomous Robots*.

Lahijanian, M., and Kwiatkowska, M. 2016. Social trust: a major challenge for the future of autonomous systems. In *AAAI Fall Symposium on Cross-Disciplinary Challenges for Autonomous Systems*, AAAI Fall Symposium. AAAI.

Lee, J. D., and See, K. A. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46(1):50–80.

Lee, D. 2016. US opens investigation into Tesla after fatal crash. *British Broadcasting Corporation (BBC) News*.

Mayer, R. C.; Davis, J. H.; and Schoorman, F. D. 1995. An integrative model of organizational trust. *Academy of management review* 20(3):709–734.

Meyer, J.-J. C.; Broersen, J.; and Herzig, A. 2014. BDI logics. In *Handbook of Epistemic Logic*. College Publications.

Schild, K. 2000. On the relationship between bdi logics and standard logics of concurrency. *Autonomous Agents and Multi-Agent Systems* 259 – 283.

Setter, T.; Gasparri, A.; and Egerstedt, M. 2016. Trust-based interactions in teams of mobile agents. In *American Control Conference*, 6158–6163.

Sweet, N.; Ahmed, N. R.; Kuter, U.; and Miller, C. 2016. Towards self-confidence in autonomous systems. In *AIAA Infotech@ Aerospace*. 1651–1652.