Cognitive Reasoning and Trust in Human-Robot Interactions[⋆]

Marta Kwiatkowska

Department of Computer Science, University of Oxford, Oxford, UK

Abstract. We are witnessing accelerating technological advances in autonomous systems, of which driverless cars and home-assistive robots are prominent examples. As mobile autonomy becomes embedded in our society, we increasingly often depend on decisions made by mobile autonomous robots and interact with them socially. Key questions that need to be asked are how to ensure safety and trust in such interactions. How do we know when to trust a robot? How much should we trust? And how much should the robots trust us? This paper will give an overview of a probabilistic logic for expressing trust between human or robotic agents such as "agent *A* has 99% trust in agent *B*'s ability or willingness to perform a task" and the role it can play in explaining trust-based decisions and agent's dependence on one another. The logic is founded on a probabilistic notion of belief, supports cognitive reasoning about goals and intentions, and admits quantitative verification via model checking, which can be used to evaluate trust in human-robot interactions. The paper concludes by summarising future challenges for modelling and verification in this important field.

1 Introduction

Autonomous robotics has made tremendous progress over the past decade, with dramatic advances in areas such as driverless cars, home assistive robots, robot-assisted surgery, and unmanned aerial vehicles. However, high-profile incidents such as the fatal Tesla crash [11] make clear the risks from improper use of this technology. Our decisions whether to rely or not on automation technology are guided by *trust*. Trust is a subjective evaluation made by one agent (the truster) about the ability or willingness of another agent (the trustee) to perform a task [6, 14]. A key aspect of a trust-based relationship is that the trustor's decision to trust is made on the expectation of benevolence from the trustee [15], and the trustor is, in fact, vulnerable to the actions of the trustee. Studies of trust in automation [12] have concluded that it is affected by factors such as reliability and predictability: it increases slowly if the system behaves as expected, but drops quickly if we experience failure. However, autonomous robots are independent decision-makers, and may therefore exhibit unpredictable, and even surprising, behaviour. Further, they need to correctly interpret the social context and form relationships with humans, thus becoming members of society.

Human relationships are built on (social) trust, which is a key influence in decisions whether an autonomous agent, be it a human or a robot, should act or not. Social trust

[⋆] This work is supported by EPSRC Mobile Autonomy Programme Grant EP/M019918/1.

is an expression of a complex cognitive process, informed by the broader context of cultural and social norms. Reasoning with trust and norms is necessary to justify and explain robots' decisions and draw inferences about accountability for failures, and hence induce meaningful communication and relationships with autonomous robots. There are dangers in acting based on inappropriate trust, for example, 'overtrust' in the Tesla crash. We need to program robots so that they can not only be trusted, but also so that they develop human-like trust in humans and other robots, and human-like relationships.

While reliability for computerised systems has been successfully addressed through formal verification techniques such as model checking, trust and ethics for robotics has only recently emerged as an area of study, in response to the rapid technological progress [10]. Elsewhere, for example in management, psychology, philosophy and economics, trust has been studied widely. Digital trust concepts are also prominent in e-commerce, where trust is based on reputation or credentials. However, the notion of trust needed for human-robot partnerships is *social trust*, which has been little studied: it involves cognitive processes (i.e. mental attitude, goals, intentions, emotion) that lead to a decision whether to trust or not, and is influenced through past experience and preferences.

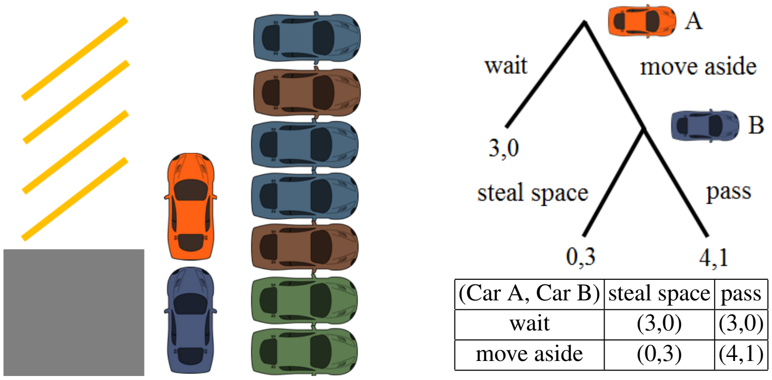
This paper gives an overview of recent progress towards a specification formalism for expressing social trust concepts. The resulting logic, Probabilistic Rational Temporal Logic (PRTL*), is interpreted over stochastic multiagent systems (essentially concurrent stochastic games) extended with goals and intentions, where stochasticity arises from randomness and environmental uncertainty. Trust is defined in terms of (*subjective*) probabilistic belief, which allows one to quantify the amount of trust as a *belief-weighted expectation*, informally understood as a degree of trust. The logic can express, for example, if A is a human rider of an autonomous car B, that "A has 99% trust in B's ability to safely reach the required destination", and "B has 90% trust in A's willingness not to give unwise instructions". The key novelty in the framework is the addition of the *cognitive* dimension, in which (human or robotic) agents carry out their deliberations prior to decision-making; once the decision has been made, the agents act on them, with the actions taking place in the usual *temporal* dimension. The logic, under certain restrictions, admits a model checking procedure, which can be employed in decision-making to evaluate and reason about trust in human-robot relationships, and to assist in establishing accountability. We illustrate the main trust concepts by means of an example, referring the reader to the details in [7].

2 An Illustrative Example

We illustrate the key features of social trust using a variant of a trust game called the Parking Game due to Vincent Conitzer [3], see Figure 2. Trust games are often used in economics, where it is assumed that players act on the basis of pure self-interest. However, experiments with human subjects consistently show that humans behave differently and are often willing to act on the assumption of the other player's goodwill.

The Parking Game illustrates a situation where cars A and B (let us assume they are autonomous) are waiting for a parking space, with car B behind A. Car A either waits or can move aside to let car B through, on the assumption that B is in a hurry and wants to pass. Car B, however, can steal A's parking space if it becomes available, or pass. Though somewhat artificial, we will also allow an iterated version of this game, where the cars return to compete for the parking space in the same order. The payoffs in this game indicate that the best outcome for both A and B is for A to move aside and B pass. As experience shows, this is a typical situation if the cars were driven by human drivers. However, according to the standard game-theoretic solution the Nash equilibrium is for A to wait, rather than move aside, to avoid the parking space being taken.

In [13] an alternative solution method is proposed that results in the equilibrium of A moving aside and B passing. A similar game is considered in [8], where the computation of the payoff is amended to include trust value. This paper puts forward a different solution, where we explicitly model the evolution of trust starting from some initial value, and update that (subjective) trust based on experience (that is, interactions between agents), preferences and context.



(Car A, Car B)	steal space	pass
wait	(3,0)	(3,0)
move aside	(0,3)	(4,1)

Fig. 1. The 'Parking Game' due to Vincent Conitzer [3] (reproduced with permission).

3 The Model

We work with stochastic multiplayer games as models, to capture both external uncertainty as well as internal probabilistic choices. We sometimes refer to players of the game as agents.

Let $\mathcal{D}(X)$ denote the set of probability distributions on a set X. For simplicity, we present a simplified variant of the model, and remark that partial observability and strategic reasoning can be handled [7].

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* $Act = \times_{A \in Ags} Act_A$ *and for each state* s *there* *exists a unique joint action $a \in Act$ such that a (non-unique) state s' is chosen with probability $T(s,a)(s')$.*

Let a_A be agent A's action in the joint action $a \in Act$. We let $Act(s) = \{a \in Act \mid T(s,a)$ 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 $s \in S$.

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.

We define a finite, resp. infinite, *path* ρ in the usual way 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 $\mathrm{FPath}^M_T(s)$ and $\mathrm{IPath}^M_T(s)$, 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 $\mathrm{last}(\rho)$ 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 $\sigma_A : \mathrm{FPath}^M_T \longrightarrow \mathcal{D}(Act_A)$ such that for all $a_A \in Act_A$ and finite path ρ it holds that $\sigma_A(\rho)(a_A) > 0$ only if $a_A \in Act_A(\mathrm{last}(\rho))$. 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 $\mathrm{Pr}^{M,\sigma}$ on $\mathrm{IPath}^M_T(s_{\mathrm{init}})$ in the standard way.

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 [2]. We work with *probabilistic* beliefs. A cognitive mechanism includes *goals*, *intentions* and *subjective preferences*. For an agent A, the idea is that, while actions Act_A represent A's actions in the *physical* space, goals and intentions 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 2. *We define a* cognitive mechanism *as a tuple $\Omega_A = (\{Goal_A\}_{A \in Ags}, \{Int_A\}_{A \in Ags}, \{gp_{A,B}\}_{A,B \in Ags}, \{ip_{A,B}\}_{A,B \in Ags})$, 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. We think of goals as abstract attitudes, for example altruism 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 refer to a stochastic multiplayer game endowed with a cognitive mechanism as an *autonomous* stochastic multiagent system. We extend the set of temporal transitions with cognitive transitions for agent A corresponding to a change of goal (respectively intention) and the transition probability function T in the obvious way. We denote by $\mathrm{FPath}^M(s)$, $\mathrm{IPath}^M(s)$, FPath^M and IPath^M the sets of paths formed by extending the sets

$\mathrm{FPath}^M_T(s)$, $\mathrm{IPath}^M_T(s)$, FPath^M_T and IPath^M_T of temporal paths with paths that interleave the cognitive and temporal transitions.

To obtain a probability measure over infinite paths $\mathrm{IPath}^M(s_{\mathrm{init}})$, we need to resolve agents' possible changes to goals or intentions. Similarly to action strategies, we define *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 $g_A : \mathrm{FPath} \longrightarrow \mathcal{D}(2^{Goal_A})$, and the intentional strategy as $i_A : \mathrm{FPath} \longrightarrow \mathcal{D}(Int_A)$. We remark that such strategies arise from cognitive architectures, with the subjective view induced by goal and intentional *preference functions*, $gp_{A,B}$ and $ip_{A,B}$, which 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. For details see [7].

Example 1. For the Parking Game example, let us consider two possible goals for A, altruism and selfishness. The intention corresponding to altruism is a strategy that always chooses to move aside, whereas for selfishness it is to choose wait. Another goal is absent-mindedness, which is associated with a strategy that chooses between moving aside and waiting at random. A preference function for B could be based on past observations that a Google car is more likely to move aside than, say, a Tesla car.

4 Probabilistic Rational Temporal Logic

We give an overview of the logic PRTL* that combines the probabilistic temporal logic PCTL* with operators for reasoning about agents' beliefs and cognitive trust. The trust operators of the logic are inspired by [4], except we express trust in terms of probabilistic 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 3. *The syntax of the language PRTL* is:*

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

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 [5, 1], 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.

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. $\mathbb{CT}_{A,B}^{\bowtie q}\psi$ 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. $\mathbb{DT}_{A,B}^{\bowtie q}\psi$ 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 Parking Game example, the formula

$$\mathbb{DT}_{A,B}^{\geq 0.7}\neg steal_A$$

where $steal_A$ is an atomic proposition, expresses that A's trust in B's willingness not to steal a space is at least 70%, and

$$\mathbb{B}_A^{\geq 0.8}\mathbb{DT}_{B,A}^{\geq 0.7} move_A$$

states that A's belief that B has at least 70% trust in its willingness to move is at least 80%, where $move_A$ is an atomic proposition. Assuming that B has absent-mindedness as its goal, and A has two goals, altruism and selfishness, with the corresponding intentions, as in Example 1, then

$$\mathbb{G}_A\neg\mathbb{DT}_{B,A}^{\geq 0.7} move_A$$

states that, for all goal changes of A, B does not trust in A's willingness to move with probability at least 70%, where $move_A$ is an atomic proposition.

We interpret formulas ϕ in an autonomous stochastic multiagent system 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. 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 corresponds to the 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 belief-weighted 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, weighted by the belief and compared to the probability bound q.

The logic PRTL* can also express strong and weak dependence trust notions of [4]. 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).

Example 3. If B is in a hurry, then

$$\mathbb{DT}^{\geq 0.9}_{B,A} \Diamond leave_B \wedge \neg \mathbb{B}^{\geq 0.9}_{B} \Diamond leave_B$$

where $leave_B$ is an atomic proposition, expresses that B's leaving the car park strongly depends on A's willingness to cooperate.

Our framework encourages collaboration by allowing agents to update their trust evaluation for other agents and to take into consideration each other's trust when taking decisions. Trust thus evolves dynamically based on agent interactions and the decision to trust can be taken when a specific trust threshold is met. Therefore our notion of social trust helps to explain cases where actual human behaviour is at variance with standard economic and rationality theories.

Example 4. For the Parking Game example, we model the evolution of trust based on interactions and prior knowledge, whereby A's trust in B decreases if B steals the space, and increases otherwise. A guards its decision whether to move aside by considering the level of trust in B's willingness not to steal, e.g. $\mathbb{DT}^{\geq 0.7}_{A,B} \neg steal_B$.

The precise value of the threshold for trust is context-dependent. The trust value higher than an appropriately calibrated level is known as 'overtrust', which can be expressed using our formalism, see [7].

5 Concluding Remarks

This paper has provided a brief overview of recent advances towards formalisation and quantitative verification of cognitive trust for stochastic multiplayer games based on [7]. Although the full logic is undecidable, we have identified decidable sublogics with reasonable complexity. As the next step we aim to implement the techniques as an extension of the PRISM probabilistic model checker [9] and evaluate them on case studies. To this end, we will define a Bellman operator and integrate with reasoning based on cognitive architectures.

This paper constitutes the first step towards developing design methodologies for capturing the social, trust-based decisions within human-robot partnerships. Pertinent scientific questions arise in the richer and challenging field of ethics and morality. How can we communicate intent in the context of human-robot interactions? How do we incentivise robots to elicit an appropriate response? How do we ensure that robotic assistants will not cause undue harm to others in order to satisfy the desires of their charge? Or that a self-driving car is able to decide between continuing on a path that will cause harm to other road-users, or executing an emergency stop which may harm passengers? These questions call for an in-depth analysis of the role of autonomous robots in society from a variety of perspectives, including philosophical and ethical, in addition to technology development, and for this analysis to inform policy makers, educators and scientists.

References

1. Andrea Bianco and Luca de Alfaro. Model checking of probabalistic and nondeterministic systems. In *FSTTCS 1995*, pages 499–513, 1995.
2. M.E. Bratman. *Intentions, Plans, and Practical Reason*. Harvard University Press, Massachusetts, 1987.
3. Vincent Conitzer, Walter Sinnott-Armstrong, Jana Schaich Borg, Yuan Deng, and Max Kramer. Moral decision making frameworks for Artificial Intelligence. In *AAAI 2017*, 2017. To appear.
4. R. Falcone and C. Castelfranchi. Social trust: A cognitive approach. In *Trust and Deception in Virtual Societies*, pages 55–90. Kluwer, 2001.
5. Hans Hansson and Bengt Jonsson. A logic for reasoning about time and reliability. *Formal aspects of computing*, 6(5):512–535, 1994.
6. Russell Hardin. *Trust and trustworthiness*. Russell Sage Foundation, 2002.
7. Xiaowei Huang and Marta Kwiatkowska. Reasoning about cognitive trust in stochastic multiagent systems. In *AAAI 2017*, 2017. To appear.
8. Benjamin Kuipers. What is trust and how can my robot get some? (presentation). In *RSS 2016 Workshop on Social Trust in Autonomous Robots*, 2016.
9. M. Kwiatkowska, G. Norman, and D. Parker. PRISM 4.0: Verification of probabilistic realtime systems. In G. Gopalakrishnan and S. Qadeer, editors, *Proc. 23rd International Conference on Computer Aided Verification (CAV'11)*, volume 6806 of *LNCS*, pages 585–591. Springer, 2011.
10. Morteza Lahijanian and Marta Kwiatkowska. 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, AAAI Press, 2016.
11. Dave Lee. US opens investigation into Tesla after fatal crash. *British Broadcasting Corporation (BBC) News*, Jul. 2016. [Online; posted 1-July-2016; http://www.bbc.co.uk/news/technology-36680043].
12. John D Lee and Katrina A See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 46(1):50–80, 2004.
13. Joshua Letchford, Vincent Conitzer, and Kamal Jain. An ”ethical” game-theoretic solution concept for two-player perfect-information games. In *Proceedings of the Fourth Workshop on Internet and Network Economics (WINE-08)*, pages 696–707, 2008.
14. Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.
15. D Harrison McKnight, Vivek Choudhury, and Charles Kacmar. Developing and validating trust measures for e-commerce: An integrative typology. *Information systems research*, 13(3):334–359, 2002.