Model of the Influence of External Signals on the Trust of the Agent in Multi Agent System

Extended Abstract

Frédérique Lalieu
Complex Cyber Infrastructure, Informatics Institute, Faculty of Science, University of Amsterdam
Amsterdam, The Netherlands
frederique.lalieu@student.uva.nl

Tomasz Zurek
Complex Cyber Infrastructure, Informatics Institute, Faculty of Science, University of Amsterdam
Amsterdam, The Netherlands
t.a.zurek@uva.nl

Tom van Engers
University of Amsterdam and TNO / Leibniz Institute
Amsterdam, The Netherlands
t.m.vanengers@uva.nl

ABSTRACT

We introduce and discuss the mechanism of the influence of external signals on the perception of benevolence, one of the components of trust of an agent in a MAS. The model presented in this paper is illustrated by a simulation experiment.

KEYWORDS

trust; trustworthiness; agent-based programming; belief revision; evidential reasoning

ACM Reference Format:
Frédérique Lalieu, Tomasz Zurek, and Tom van Engers. 2025. Model of the Influence of External Signals on the Trust of the Agent in Multi Agent System: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Trust is a fundamental factor in establishment and maintenance of interpersonal relations and relationships between humans and institutions [3]. The main objective of our project is to create a model of the dynamics of trustworthiness in MAS agents in response to external signals about these agents and their actions. This effort consisted of two parts: first, the development of a model for updating the belief base in response to external signals; second, the development of a model for updating the evaluation of the components of trust based on the updated belief base. Here we focus on a comprehensive analysis of one of the components of trust and prepare the grounds for the analysis of its other elements. Usually (e.g. [2, 5]), trustworthiness is defined by its three components: competence, benevolence, and integrity. In this work we focus on the modeling of the dynamics of benevolence, because it is the most complex and difficult to represent element of trustworthiness.

2 THE COMPLEX NATURE OF TRUSTWORTHINESS

For the definition of trustworthiness, we rely on the analysis of [1]. They strictly distinguish between trust and trustworthiness: trust is a property of the trustor (towards a trustee), while trustworthiness is a property of the trustee. Trust is built on the basis of trustworthiness, personality of the trustor, the plausibility gap (presence or lack of evidence), etc. In our model, the trustworthiness of a trustee is the competence, benevolence and integrity of the trustee evaluated by the trustor. Combining the definitions of both [5, 6], the essence of benevolence is to do good things for the trustor, even though it is not necessarily beneficial to the trustee.

3 THE MODEL

Static model. Our model of the concept of benevolence will be constructed on the basis of some concepts of value-based reasoning from [10]. We are going to use *values* as the central concept allowing for representation of the agents' goals. The key assumption is that every goal, understood as a particular state of affairs to be reached, satisfies (promotes or demotes) some values to a certain extent. Therefore, the comparison between goals will be based on the levels of satisfaction of values. The initial version of the static model of benevolence was introduced in [4, 9]. Below we present its improved version.

- Let $V = \{v_a, v_b, v_c, ...\}$ be a set of values and $A = \{a_T, a_p, a_q, ...\}$ be a set of agents. Suppose that agent a_T is a trustor and a_p and a_q are trustees.
- Let $P_{a_p} = \{SP^1_{a_p}, SP^2_{a_p}, ..., SP^n_{a_p}\}$ be the plan of agent a_T to be executed by Trustee a_p consisting of a set of subplans. By P we denote a set of all plans.
- Let $g^{SP}_{a_T}$ be a proposition representing an atomic goal of a particular subplan SP of agent a_T in a particular moment of time. By $G^P_{a_T}$ we denote a set of goals in plan P. G denotes a set of all goals[1].
- Let $\Phi_v : A \times 2^G \longrightarrow \langle 0; 1)$ be a function returning the level of satisfaction of value v by a subset of G in the eyes of agent a. Let $\Phi = \{\Phi_{v_a}, \Phi_{v_b}, ...\}$ be a set of all functions Φ_v.

Note that the agent (trustee) may have a different attitude towards values, and the trustee's willingness to demote their goals to support the trustor's ones, is the possibility to demote a set of values, each of which can have a different threshold:

Definition 1. *Let* $\Gamma : A \times V \longrightarrow \langle 0; 1)$ *be a function representing benevolence for every agent. It returns the maximal acceptable levels of the demotion of values' from set* V *w.r.t initial goal of an agent* $a \in A$

[1] Note that one plan may satisfy multiple goals and different plans may achieve the same goals, perhaps at different costs

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA.

By $\Gamma_{a_T}(a_p, v_m)$ we denote the evaluation of benevolence of agent a_p w.r.t. value v_m made by agent a_T.

A trustor assumes that a trustee can accept a new goal only if for every value, a new goal does not demote the value to a higher extent than the benevolence level allows.

Definition 2. *Let $G_{a_p}^{P_{a_p}}$ be a set of goals of an initial plan of agent a_p. A new plan P'_{a_p} will be acceptable for agent a_p and agent a_p will be sufficiently benevolent for adopting this plan in the view of trustor a_T, if: $m\forall_{v_x \in V}(\Phi_{v_x}(a_p, G_{a_q}^{P_{a_p}}) < (\Gamma_{a_T}(a_p, v_x) + \Phi_{v_x}(a_p, G_{a_p}^{P'_{a_p}})))$ By $BEN(a_p, P'_{a_p})$ we denote that agent a_p is sufficiently benevolent for performing plan P'_{a_p} .*

The model introduced in this section allows for finding which of the potential trustees are sufficiently benevolent to fulfill the delegated task.

Trustworthiness dynamics. For the sake of this work, as an external signal we treat any information the agent receives from the environment. For the sake of simplicity, we do not discuss the topic of behavior consistency and source selection in this paper (but we are aware of it), but rather assume that the process of selection is done and that we have two types of signals (messages):

- $SUCC_{a_p}$ is a one-element set containing the recent plan of which the trustor knows that it has been successfully performed by agent a_p.
- $FAIL_{a_p}$ is a one-element set containing the plan of which the trustor knows that it has not been successfully performed by agent a_p. For the sake of simplicity we assume that set $FAIL_{a_p}$ contains plans that failed because of the lack of benevolence.

Dynamic model. An agent a_T updates their estimation of the benevolence of another agent a_p when they believe that a_p accepted or refused an offer to switch to another plan. Let's consider a scenario in which trustor a_T believes that trustee a_p recently accepted new plan $plan_N$ (a new plan $plan_N$ has been added to $SUCC(a_p)$) which replaced a_p's original plan $planI$. The trustor also has beliefs about the levels of satisfaction of all values by the outcomes of a_p's initial ($plan_I$) and the new plan ($plan_N$). For value v_x and agent a_p, the old evaluation of the lower boundary of benevolence we denote by $\Gamma_{a_T}^{\downarrow}(a_p, v_x)$, and the new evaluation by $\Gamma_{a_T}^{\downarrow}(a_p, v_x)'$:

Definition 3. *If a_T is a trustor, a_p a trustee, then a_T calculates a new lower boundary of a_p's benevolence on the basis of the formula: $\forall_{v_x \in V}((\Gamma_{a_T}^{\downarrow}(a_q, v_x)') = max(\Gamma_{a_T}^{\downarrow}(a_q, v_x), (\phi_{v_x}(a_p, G_{a_p}^{planI})- (\phi_{v_x}(a_p, G_{a_p}^{planN})))$*

In simple terms, the agent is perceived to be as least as benevolent as his/her previous successfully fulfilled plans. Conversely, if agent a_T notices that agent a_p refused to switch from $plan_I$ to $plan_N$ (planN is added to $FAIL_{a_p}$), he can revise the upper boundary of the benevolence interval with respect to some values. The evaluating agent should know on the basis of which values the offer was rejected. Let $V_D \subseteq V$ be the set of values that were decisive in a_p's decision to refuse the offer[2]. By $\Gamma_{a_T}^{\uparrow}(a_p, v_x)$ we denote the old evaluation of the upper boundary of benevolence, while a new one we denote by $\Gamma_{a_T}^{\uparrow}(a_p, v_x)'$:

Definition 4. *If a_T is a trustor, a_p a trustee, then a_T can calculate a new upper boundary of a_p's benevolence on the basis of formula: $\forall_{v_x \in V_D}((\Gamma_{a_T}^{\uparrow}(a_q, v_x)') = min(\Gamma_{a_T}^{\uparrow}(a_q, v_x), (\phi_{v_x}(a_p, G_{a_p}^{planI})- (\phi_{v_x}(a_p, G_{a_p}^{planN})))$*

The upper boundary should be exclusive because the agent will refuse to make a sacrifice if he has refused the same sacrifice before. Note that there is a gray area between those the lower and the upper boundary. There is, obviously, no simple solution on how to evaluate plans located in this area, because people differ in the way they deal with uncertainty. To represent this mechanism we assume a set of functions Ψ:

Definition 5. *Let $\Psi = \{\Psi_{a_T}, \Psi_{a_p}, \Psi_{a_q}, ...\}$, where $a_T, a_p, p_q \in A$, be a set functions $\Psi_a : [0..1] \times [0..1] \longrightarrow [0..1]$, which on the basis of upper and lower bound of perceived benevolence calculates a joint evaluation of potential trustee's benevolence.*

Note that this function is agent-dependent: every agent may have a different way of estimation of someone else's benevolence. For the sake of this work we assume that a new level of benevolence of an agent is a mean of a new upper and lower boundary of benevolence: $\Gamma_{a_T}(a_q, v_x)' = \Psi_{A_T}(\Gamma_{a_T}^{\uparrow}(a_q, v_x)', \Gamma_{a_T}^{\downarrow}(a_q, v_x)') = (\Gamma_{a_T}^{\uparrow}(a_q, v_x)' + \Gamma_{a_T}^{\downarrow}(a_q, v_x)')/2$

Implementation. The proof of concept of our model will be tested on the *belief-desire-intention* (BDI) model [7] based framework and language ASC2. The model implementation can be found on GitHub[3]

4 DISCUSSION AND CONCLUSIONS

In most of the existing models (e.g [2, 8]) benevolence is represented by a number. Our first observation is that benevolence is more complex and should be presented in the light of the agent's (trustee's) goal: how much they can sacrifice with respect to their initial plans. In order to do that, we should introduce the notion of goals. On the basis of that, we can also observe that the agent may have varying willingness to sacrifice different values. Moreover, we observed that the trustor, while evaluating a potential trustee, may receive two types of signals: positive and negative, impacting respectively the lower and upper boundary of the benevolence interval. Between them, there is a gray area in which a trustor should estimate the potential trustee's benevolence.

In future works, we plan to perform experiments testing whether it is possible to observe the trust-related phenomena observable in social sciences in the sociotechnical environment of artificial agents. In a longer run, our project will help to create a theoretical background for the creation of trustworthy and trusting autonomous devices.

[2] Although in many situations a trustor does not know which values were decisive, for the sake of simplicity, we assume that they are known.

[3] https://github.com/FrederiqueLalieu/Trust-dynamics-benevolence

REFERENCES

[1] C. Castelfranchi and R. Falcone. 2010. *Trust theory : a socio-cognitive and computational model.* John Wiley & Sons Ltd., UK.

[2] A. Delijoo. 2021. *Computational trust models for collaborative network orchestration.* Ph.D. Dissertation. University of AMsterdam.

[3] M. Firmansyah, Riski Amelia, Rizky Jamil, Faturochman Faturochman, and Wenty Minza. 2019. BENEVOLENCE, COMPETENCY, AND INTEGRITY: WHICH IS MORE INFLUENTIAL ON TRUST IN FRIENDSHIPS? *Jurnal Psikologi* 18 (08 2019), 91–105. https://doi.org/10.14710/jp.18.1.91-105

[4] Basten Leeftink, Britta Abbink Spain, Tomasz Zurek, and Tom van Engers. [n.d.]. A Computational Model of Trustworthiness: Trust-based interactions between agents in Multi Agent System. Proceedings of 17th Internaitonal Conference on Agents and Artificial Intelligence. To appear in 2025.

[5] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. http://www.jstor.org/stable/258792

[6] D. Harrison Mcknight, Michelle Carter, Jason Bennett Thatcher, and Paul F. Clay. 2011. Trust in a Specific Technology: An Investigation of Its Components and Measures. *ACM Trans. Manage. Inf. Syst.* 2, 2, Article 12 (jul 2011), 25 pages. https://doi.org/10.1145/1985347.1985353

[7] Anand S. Rao and Michael P. Georgeff. 1995. BDI Agents: From Theory to Practice. In *Proceedings of the First International Conference On Multi-Agent Systems (ICMAS-95).* 312–319.

[8] Alessandro Sapienza, Filippo Cantucci, and Rino Falcone. 2022. Modeling Interaction in Human–Machine Systems: A Trust and Trustworthiness Approach. *Automation* 3, 2 (2022), 242–257. https://doi.org/10.3390/automation3020012

[9] Adam Wyner, Tomasz Zurek, and Tom van Engers. 2024. The model of benevolence for trust in Multi-Agent System. Agents and Multi-agent Systems: Technologies and Applications 2024, Proceedings of 18th KES International Conference, KES-AMSTA 2024, June 2024, Springer, to appear in March 2025.

[10] Tomasz Zurek. 2017. Goals, values, and reasoning. *Expert Systems with Applications* 71 (2017), 442 – 456. https://doi.org/10.1016/j.eswa.2016.11.008