Actual Trust in Multiagent Systems

Extended Abstract

Michael Akintunde
King's College London
London, United Kingdom
michael.akintunde@kcl.ac.uk

Vahid Yazdanpanah
University of Southampton
Southampton, United Kingdom
v.yazdanpanah@soton.ac.uk

Asieh Salehi Fathabadi
University of Southampton
Southampton, United Kingdom
a.salehi-fathabadi@soton.ac.uk

Corina Cirstea
University of Southampton
Southampton, United Kingdom
cc2@ecs.soton.ac.uk

Mehdi Dastani
Utrecht University
Utrecht, Netherlands
m.m.dastani@uu.nl

Luc Moreau
King's College London
London, United Kingdom
luc.moreau@kcl.ac.uk

ABSTRACT

We study how trust can be established in multiagent systems where human and AI agents collaborate. We propose a computational notion of *actual trust*, emphasising the modelling of trust based on agents' capacity to deliver tasks in prospect. Unlike reputation-based trust, we consider the specific setting in which agents interact and model a forward-looking notion of trust. We provide a conceptual analysis of actual trust's characteristics and highlight relevant trust verification tools. By advancing the understanding and verification of trust in collaborative systems, we contribute to responsible and trustworthy human-AI interactions, enhancing reliability in various domains.

KEYWORDS

Trust; Multiagent Systems; Human-AI Interactions

ACM Reference Format:
Michael Akintunde, Vahid Yazdanpanah, Asieh Salehi Fathabadi, Corina Cirstea, Mehdi Dastani, and Luc Moreau. 2024. Actual Trust in Multiagent Systems: Extended Abstract. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 3 pages.

1 INTRODUCTION

In the context of responsible and trustworthy human-AI collaboration [22], the need for computational tools and methods to verify *actual trust* between system components, in terms of their capacity to deliver tasks, is paramount. This research emphasises the significance of establishing trust in multiagent systems (MAS), where human and AI agents collaborate to achieve shared tasks. We propose a novel perspective on trust, wherein an agent or group, referred to as α, is considered trusted by another agent or group, referred to as β, with respect to a task T, if β can verify that α has the necessary strategic ability and epistemic capacity to successfully accomplish T, and that α has the intention to accomplish T. This view on trust in MAS differs to trust solely based on an agent's reputation or statistical analysis of their historical behaviour; it emphasises the importance of considering the actual setting in which agents interact; our approach underscores the significance of verifying a group of agents' true capability to deliver in the current context.

Inspired by Halpern [12], we advocate for distinguishing history-based retrospective reasoning from prospective reasoning about what agents can actually ensure in a given setting, integrating formal logic-based methods within the framework of a MAS. In particular, we distinguish between what is *typically* delivered by agents and what agents *actually* (i.e. they have the ability and intention to) deliver, and hence are trusted for in a given setting. To that end, we argue that true trust verification necessitates an assessment of what agents are genuinely capable of accomplishing. Therefore, we propose a computational notion of *actual trust*.

Trust is a multifaceted concept as it encompasses the interplay between an agent's ability, knowledge, and exhibits a temporal dimension [20]. In the literature on trust in MAS, we observe three dominant perspectives; trust can be i) modelled based on the *cognitive states* [9] of agents [13, 14], ii) *reputation-based* [21], focusing on past behaviour and reputation of agents, or iii) assumed as bidirectional *relations* given at design-time [3, 5–7], which may have limited applicability in dynamic environments. Humans tend to place unwarranted levels of trust in AI systems during their interactions [19, 23], which underscores the need of methods for verifying the trustworthiness of an agent in a particular context, rather than making generalised assumptions based on past interactions.

We distinguish between two types of trust: *retrospective trust* that reasons about trusting an agent based on the past and *prospective trust* which looks at the abilities of agents and what they can deliver in the future. We denote the former as typical trust and the latter as actual trust. In comparison to reputation-based methods with a retrospective approach to trust [4, 20], we maintain a prospective view of trust and build trust based on agents' ability, their knowledge of the environment and what they intend to achieve in a MAS.

2 REASONING ABOUT ACTUAL TRUST

Trust combines strategic, epistemic and social concepts. To mathematically reason about trust in a MAS, we ground our semantics on an *interpreted system* (IS) [8]. We express our notion of trust in terms of ATLK formulae [18], aiming to enable a direct transformation from the trust verification problem into an ATLK model-checking problem. ATLK [18] combines Alternating-time Temporal Logic [2]

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

(a generalisation of CTL [15]) with modal operators to reason about the knowledge of agents in a MAS. Here we focus on a necessary fragment to define trust modalities consisting of those given in "Vanilla ATL" with the knowledge operator $K_i\varphi$, "agent i knows φ".

We are modelling *trust under perfect information*. That is, what a group intends to do is known among the group members, so due to the public declaration of intentions, what a group intends to deliver is in a sense also what every individual within the group intends to do as well. We define an *interpreted system with intentions* (ISI) as an extension of IS to have each agent i associated with a consistent set of *intentions* $\mathcal{I}_i \subseteq 2^{\Phi} \setminus \emptyset$, a finite set of k propositions $\{\varphi_1, \ldots, \varphi_k\}$, with each $\varphi \in \Phi$ being propositional formulae, that i intends to bring about irrespective of the global state of the system and irrespective of all strategies of any agent in the system. We assume a consistency constraint on each $\mathcal{I}_i$; we do not allow the intention set $\mathcal{I}_i$ for agent i to consist of both p and $\neg p$ for any proposition p. An IS is a special case of an ISI where all agents intend every possible goal. Unlike in [16], intentions here are not bound to states or strategies; intending to bring about one or more propositions is orthogonal to the agent's ability to do so.

We assume the specification language $\mathcal{L}$, containing the standard Boolean connectives of CTL. In terms of the trustee β, a group of potentially trusted agents α and task T (see Section 1), we take an agent i as β, the group of agents Γ as α and our task T as the formula φ. We assume the *trust operator* $\mathcal{T}$ which takes as input an agent i, a group of agents Γ and an $\mathcal{L}$ formula φ. The formula $\mathcal{T}_i(\Gamma, \varphi)$ is read as "agent i trusts Γ to bring about φ". Formally:

Definition 1 (I trust Γ if I know they can deliver). *Given a Kripke model associated with an interpreted system with intentions $\mathcal{M}_{ISI}$, we say that $(\mathcal{M}_{ISI}, q^0) \models \mathcal{T}_i(\Gamma, \varphi)$ iff for all $q^K \in Q$ we have that if $q^0 \sim_i q^K$ then there exists a (collective) strategy s_Γ for Γ, and action $a_\Gamma \in s_\Gamma(q^K_\Gamma)$ such that for all states q^1 such that $q^K \rightarrow_a q^1$, we have that $\varphi \cap \bigcap_{i\in\Gamma} \mathcal{I}_i$ is consistent and $(\mathcal{M}_{ISI}, q^1) \models \varphi$.*

That is, φ is consistent with each agent's intentions. Here, Q is the set of *ISI*'s reachable global states and $\sim_i$ is the epistemic indistinguishability relation [8, p. 117] for i. Trust here is defined in terms of what agents intend to deliver regardless of their ability to deliver; one may intend φ regardless of its ability to deliver it from any local state. The intersection $\cap_{i\in\Gamma}\mathcal{I}_i$ finds a consistent set of goals that all agents intend to deliver. It is permitted for $i \in \Gamma$ or $\Gamma = \{i\}$; agent i trusts that it can cooperate with Γ to bring about φ, and that i has trust in itself that it can bring about φ respectively, regardless of what $Agt \setminus \Gamma$ does, where Γ is the global set of agents in the system. We highlight interesting properties such as non-monotonicity; one can check that when considering intentions, if $\Gamma \subseteq \Gamma'$, the agent i trusting Γ for ϕ does not necessarily imply that i trusts Γ' for ϕ.

The Bit Transmission Problem (BTP) [8, p. 114] can be modelled as an ISI. Assume a corresponding ISI, *BISI* such that all agents (sender S, receiver R and environment E) intend for acknowledgements to always be received, i.e. $\mathcal{I}_S = \mathcal{I}_R = \mathcal{I}_E =$ **recack**, where **recack** is an atomic proposition representing all global states where R has the bit value and S has the acknowledgement. One can check whether $\mathcal{M}_{BISI}, q \models \mathcal{T}_S(R, \textbf{recack})$ for $q \in I$, i.e. the sender trusts the receiver in bringing about **recack**, where I is the set of initial global states where R has yet to have been sent the bit.

3 DISCUSSION: EXPRESSIVITY FOR MODELLING TRUST DYNAMICS

Trust is Bounded by Knowledge. Actual trust is limited by an agent's knowledge; an agent's trust in other agents is dependent on the information it possesses and its ability to discern and evaluate the ability of others. We account for the relationship among states that an agent may not be able to differentiate due to its limited knowledge. For $\mathcal{T}_i(\Gamma, \varphi)$ to hold, the trustee must have sufficient information to assess the potential consequences of the trusted agents' actions and anticipate the states they will reach as a result. The trustor(s) must possess the necessary knowledge for fulfilling a given task. We capture the epistemic dynamics of trust and applicability for reasoning about trust in real-world scenarios.[1]

Trusting Coalitions. The relationship between individual- and collective-level trust is rooted in ATL and the semantic machinery that we used to model trust as it allows us to reason about collective-level capacities, knowledge of agent groups, and accordingly our notion of actual trust in MAS. Our notion is expressive enough to evaluate if for an agent i trusting agent j regarding a task T, whether it is reasonable to also trust any group J including j for delivering T. This requires considering whether their intentions are aligned on top of their strategic ability to deliver the task in question. Trust in an individual may not necessarily extend to encompass trust in larger groups including that individual. Our notion of trust allows for reasoning about the expansion of trust beyond the individual level, enabling us to consider trust dynamics within collective entities. By recognising such relationships between individual and collective trust, we gain a better understanding of trust dynamics in human-AI systems.

Fine-tuning Trust. We take into account the localised nature of trust within a specific situation; here trust is state-dependent. An agent i trusting agent j for task T in state q does not necessarily imply that i also trusted j in previous states through the history of states that ends in q. The key here is that we allow for fine-tuning and updating of trust; it can be adjusted and refined based on the current state and the dynamics of the situation. By incorporating this flexible understanding of trust into our model, we enable the ability to model and reason about trust in a dynamic and adaptable manner.

4 FUTURE CONTRIBUTIONS

We wish to explore different notions of trust, support multistep strategies, and eventually curate a framework for reasoning about trust, allowing also for quantification [24]. We will also utilise Event-B [1, 11, 17] to explore *refinement-based* [10] formal modelling and verification techniques for actual trust.

ACKNOWLEDGMENTS

This work is supported by EPSRC through the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1), a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems, and the platform grant entitled "AutoTrust: Designing a Human-Centered Trusted, Secure, Intelligent and Usable Internet of Vehicles" (EP/R029563/1).

[1]We highlight that as we modelled our notions in ATL, verifying actual trust can be implemented in standard model-checking tools such as MCMAS [18].

REFERENCES

[1] J-R. Abrial. 2010. *Modeling in Event-B: System and Software Engineering*. Cambridge University Press, Cambridge, UK.

[2] R. Alur, T. A. Henzinger, and O. Kupferman. 2002. Alternating-Time Temporal Logic. *J. ACM* 49, 5 (2002), 672–713.

[3] J. Bentahar, N. Drawel, and A. Sadiki. 2022. Quantitative Group Trust: A Two-Stage Verification Approach. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS'22)*. International Foundation for Autonomous Agents and Multiagent Systems, Auckland, New Zealand, 100–108.

[4] Chris Burnett, Timothy J Norman, and Katia Sycara. 2011. Trust decision-making in multi-agent systems. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11)*. AAAI Press, Barcelona, Catalonia, Spain, 115–120.

[5] Nagat Drawel, Jamal Bentahar, Amine Laarej, and Gaith Rjoub. 2022. Formal verification of group and propagated trust in multi-agent systems. *Autonomous Agents and Multi-Agent Systems* 36, 1 (2022), 19.

[6] N. Drawel, J. Bentahar, and E. Shakshuki. 2017. Reasoning about Trust and Time in a System of Agents. *Procedia Computer Science* 109 (12 2017), 632–639.

[7] N. Drawel, A. Laarej, J. Bentahar, and M. El Menshawy. 2022. Transformation-based model checking temporal trust in multi-agent systems. *Journal of Systems and Software* 192 (2022), 111383.

[8] R. Fagin, J. Y. Halpern, Y. Moses, and M. Y. Vardi. 1995. *Reasoning about Knowledge*. MIT Press, Cambridge.

[9] R. Falcone and C. Castelfranchi. 2001. Social Trust: A Cognitive Approach. In *In Trust and Deception in Virtual Societies*. Springer, Berlin, 55–90.

[10] Asieh Salehi Fathabadi and Vahid Yazdanpanah. 2023. Trust modelling and verification using Event-B. In *Proceedings of the Fifth Workshop on Formal Methods for Autonomous Systems (FMAS'23)*. EPTCS, Leiden, Netherlands, 10–16.

[11] Hang-Jiang Gao, Zheng Qin, Lei Lu, Li-Ping Shao, and Xing-Chen Heng. 2007. Formal specification and proof of multi-agent applications using event b. *Information Technology Journal* 6, 7 (2007), 1181–1189.

[12] Joseph Y Halpern. 2016. *Actual causality*. MIT Press, Cambridge, Massachusetts, United States.

[13] X. Huang and M. Kwiatkowska. 2017. Reasoning about Cognitive Trust in Stochastic Multiagent Systems. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*. AAAI Press, San Francisco, California, USA, 3768–3774.

[14] Xiaowei Huang, Marta Kwiatkowska, and Maciej Olejnik. 2019. Reasoning about cognitive trust in stochastic multiagent systems. *ACM Transactions on Computational Logic (TOCL)* 20, 4 (2019), 1–64.

[15] Michael Huth and Mark Ryan. 2004. *Logic in Computer Science: Modelling and reasoning about systems*. Cambridge University Press, Cambridge, United Kingdom.

[16] Wojciech Jamroga, Wiebe van der Hoek, and Michael Wooldridge. 2005. Intentions and strategies in game-like scenarios. In *Progress in Artificial Intelligence: 12th Portuguese Conference on Artificial Intelligence, EPIA'05*. Springer, Covilhã, Portugal, 512–523.

[17] Arnaud Lanoix. 2008. Event-B Specification of a Situated Multi-Agent System: Study of a Platoon of Vehicles. In *Proceedings of the Second IEEE/IFIP International Symposium on Theoretical Aspects of Software Engineering, (TASE'08)*. IEEE Computer Society, Nanjing, China, 297–304.

[18] A. Lomuscio, H. Qu, and F. Raimondi. 2017. MCMAS: A Model Checker for the Verification of Multi-Agent Systems. *Software Tools for Technology Transfer* 19, 1 (2017), 9–30.

[19] Mohammad Reza Mousavi, Ana Cavalcanti, Michael Fisher, Louise Dennis, Rob Hierons, Bilal Kaddouh, Effie Lai-Chong Law, Rob Richardson, Jan Oliver Ringer, Ivan Tyukin, et al. 2023. Trustworthy Autonomous Systems Through Verifiability. *Computer* 56, 2 (2023), 40–47.

[20] Sarvapali D Ramchurn, Dong Huynh, and Nicholas R Jennings. 2004. Trust in multi-agent systems. *The knowledge engineering review* 19, 1 (2004), 1–25.

[21] Sarvapali D Ramchurn, Nicholas R Jennings, Carles Sierra, and Lluis Godo. 2004. Devising a trust model for multi-agent interactions using confidence and reputation. *Applied Artificial Intelligence* 18, 9-10 (2004), 833–852.

[22] Sarvapali D Ramchurn, Sebastian Stein, and Nicholas R Jennings. 2021. Trustworthy human-AI partnerships. *Iscience* 24, 8 (2021), 102891.

[23] Paul Robinette, Wenchen Li, Robert Allen, Ayanna M Howard, and Alan R Wagner. 2016. Overtrust of robots in emergency evacuation scenarios. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. IEEE, Christchurch, New Zealand, 101–108.

[24] Vahid Yazdanpanah and Mehdi Dastani. 2016. Quantified degrees of group responsibility. In *Coordination, Organizations, Institutions, and Norms in Agent Systems*. Springer, Cham, 418–436.