HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240179

Formal Specification of Actual Trust in Multiagent Systems

Michael AKINTUNDE [a,1], Vahid YAZDANPANAH [b], Asieh SALEHI FATHABADI [b], Corina CIRSTEA [b], Mehdi DASTANI [c] and Luc MOREAU [a]

[a] *King's College London, London, United Kingdom*
[b] *University of Southampton, Southampton, United Kingdom*
[c] *Utrecht University, Utrecht, The Netherlands*

ORCiD ID: Michael Akintunde https://orcid.org/0000-0002-5031-8813, Vahid Yazdanpanah https://orcid.org/0000-0002-4468-6193, Asieh Salehi Fathabadi https://orcid.org/0000-0002-0508-3066, Corina Cirstea https://orcid.org/0000-0003-3165-5678, Mehdi Dastani https://orcid.org/0000-0002-4641-4087, Luc Moreau https://orcid.org/0000-0002-3494-120X

Abstract. This research focuses on establishing trust in multiagent systems where human and AI agents collaborate. We propose a computational notion of *actual trust*, emphasising the modelling of an agent's capacity to deliver tasks. Unlike reputation-based trust or performing a statistical analysis on past behaviour, our approach considers the specific setting in which agents interact. We integrate non-deterministic semantics for capturing inherent uncertainties within the behaviour of a multiagent system, but stress the importance of verifying an agent's actual capabilities. 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, this research contributes to responsible and trustworthy human-AI interactions, enhancing reliability in various domains.

Keywords. Trust, Multiagent Systems, Human-AI Interactions

1. Introduction

We are seeing a rapid adoption of AI agents being used for safety-critical tasks in the real-world and interacting with humans. A crucial step towards having symbiotic, responsible and trustworthy human-AI partnerships [1] is through the development of computational tools and methods to reason about how different components of such systems trust each other. We focus on *actual trust*[2], defined in terms of agents' capacity to deliver tasks. Specifically, 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 a *trustee* agent or group, referred to as β,

[1] Corresponding Author: Michael Akintunde, michael.akintunde@kcl.ac.uk
[2] The essence of ideas presented here are explored in [2].

is considered trusted by another *trustor* agent or group, referred to as α, with respect to a specific task, denoted as T, if α can verify that β has the necessary strategic ability and epistemic capacity to successfully accomplish T, and β has the intention to accomplish T. This view on trust in MAS complements modelling trust solely based on an agent's reputation or through a statistical analyses of historical behaviour. Instead, it emphasises the importance of considering the actual setting in which agents interact. Although statistical methods can be employed to narrow down the list of trusted agents for a given task, our approach underscores the significance of verifying a collective's true capability to deliver in the current context.

Drawing inspiration from Halpern [3], 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 make a distinction between what is *typically* delivered by agents and what agents *actually* (i.e. they have the ability and intention to) deliver, and hence trusted for in a given setting (which can be verified based on the agents' available actions and how such actions affect the system at hand and its properties). To that end, we argue that true trust verification necessitates an assessment of what agents are genuinely capable of accomplishing. Consider an autonomous delivery vehicle (ADV) tasked with transporting goods. Even if it was successful in former deliveries, it may currently have a low battery and is unable to achieve tasks, so the level of trust one has for the ADV needs adjustments based its current situation, regardless of its previous capabilities. We propose a computational notion of *actual trust* in MAS, which encompasses a comprehensive conceptual analysis of its key characteristics. We highlight the relevance of various formal verification tools that can be employed in MAS to ensure the delivery of trustworthy outcomes. We relax the strong assumption common in modelling MAS that agents have full observability; our trust model captures real-life uncertainties by assuming imperfect information.

This contributes to the ongoing exploration of trust in MAS by providing a robust computational foundation for assessing and verifying trust between human and AI agents. By adopting this perspective, we anticipate advancements in the development of responsible and trustworthy collaborative systems, paving the way for more effective and reliable human-AI interactions in various domains.

2. Related Work

We remark that while the work of [4,5,6,7] assumes the existence of trust relations at design-time, and allows for reasoning about more complex trust dynamics on top of the trust relations they assume, we formulate trust as a notion which emerges through the dynamic evolutions of the multi-agent system. Their trust relation is static but we allow modelling and reasoning about trust in given worlds (global states) of the MAS. Differently also to their work, we consider intentions, although the set of intentions that we consider for our trust semantics is static. Following Cohen and Levesque [8], we refer to the set of intentions as goals that an agent or agent group has chosen and is committed to delivering. In the rest of the text, we may refer to the elements of this set as intended goals or simply as goals. In comparison to reputation-based methods with a retrospective approach to trust [9,10], 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. Different to our focus on trust, a complementary line of work is the development of methods to formalise notions of responsibility in groups such as in [11], where strategic, probabilistic and temporal modalities are used to reason about a group's responsibility for taking risks. We also remark in comparison to the more abstract logic of [12], our focus here is to ground the logic on a suitable computational model and to be able to apply automated verification techniques to reason about trust in a computationally feasible manner. We do not consider beliefs in the sense of [13] in our modelling of trust.

2.1. Conceptual Analysis: Trust as a Prospective Concept

Trust between groups of agents is inherently a social phenomenon, with multiple defining characteristics such an agent's ability and knowledge, which also exhibits a temporal dimension [10]. Trust in MAS has been widely studied in the literature, which we find centers on three dominant themes:

Cognitive Trust Modelling. One influential perspective on trust is presented in [14], where trust is modelled based on the cognitive states of agents. Through this, an agent can be trusted if their beliefs and intentions, and accordingly their plan of action is aligned with our intended plans and intended future. However, in complex systems like human-AI interactions, determining an agent's true intentions can be challenging, and intention elicitation remains an open problem in AI systems. While cognitive models of trust offer a high-level understanding of trust dynamics, their implementation in large-scale human-AI settings, such as smart mobility applications, requires further research.

Reputation-Based Trust. Another approach to trust, proposed by [15], focuses on past behaviour and agents' reputation. It suggests that agents who have consistently performed according to plans in the past can be trusted to deliver similar performance in the future. This is particularly applicable in domains with predictable environments, such as closed world databases where agents follow safe protocols for data updates. However, when considering trust for a specific task, it is more reliable to evaluate an agent's available actions and knowledge in the current state of the MAS. What agents delivered in the past can be used as a means to limit the search space for trusted agents to deliver a particular task but may result in biased evaluations if used as the sole measure for establishing trust.

From Trust Relations to Collective Trust. More recent lines of research on trust [5] assumes trust relations and builds a framework that analyses how these assumed trust relations cascade in the system, leading to the formation of collective-level trust among coalitions. However, the core assumption of this framework is access to a social network of bidirectional trust relations. Although in some settings such a social network may be available and fixed, the assumption of fixed trust relations as given may limit the applicability of this approach in dynamic environments. More research in this line shows how different types of knowledge [16] and uncertainties regarding the epistemic state of agents and agent groups can be integrated into modelling trust dynamics.

Highlighted by [17] is the phenomenon of humans tending to place unwarranted levels of trust in AI systems during their interactions. This has been observed in studies of humans interacting with robots in high-risk scenarios [18]; humans can "overtrust"

robots when they are observed to malfunction during prior interactions. This underscores the need for the development of methods aimed at verifying the trustworthiness of a specific AI agent within a particular context, rather than making generalised assumptions based solely on past interactions. This holds significant promise in refining the trustworthiness of AI systems and mitigating instances of unjustified trust solely rooted in historical interactions.

Against this background, we believe it is crucial to distinguish between two types of trust: retrospective trust that reasons about trusting an agent based on 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 (as it relies on the typical, historical, and statistical data on agent's past performance and reputation) and the latter as *actual* trust as it relies on the actual state of the system and what the agent can deliver under their strategic and epistemic limitations. We ideate that trusting an agent solely based on historical data disregards the contextual factors and the specific requirements of the current scenario. Instead, a more comprehensive approach to establishing trust is needed, which is not solely based on the performance of an agent's past behaviour but also relies on an agent's ability to deliver a task in the future, determined by the current system state. This can be modelled as a concurrent game, where trust in a group of agents with respect to a specific task can be determined based on the foreseeable consequence of joint actions. An intelligent agent can utilise both actual and retrospective trust notions to obtain a comprehensive evaluation of who to trust.

As a motivating example, The Bit Transmission Problem [19] underscores the limitations associated with relying solely on an agent's past behaviour to establish trust. This problem involves the transmission of the value of a bit from a (potentially human) source agent to a destination agent. Traditional trust models, which hinge on reputation-assessments and past behaviour, would suggest that an agent with a proven track record of successfully communicating information can be considered trusted for future transmissions. However, this approach fails to consider the dynamic nature of the system and the specific contextual factors at play[3]. It is crucial to evaluate the agents' current capacities (as actions they possess and how their actions may affect the environment), knowledge (as what they know about their own ability and can distinguish in possible next states of the system), intentions of the agents, and the specific requirements of the task within the existing system state. By examining the agents' available actions and knowledge within the present system configuration, a more informed judgement can be made regarding their reliability in transmitting bits. Therefore, a shift in focus is necessary, moving away from relying solely on historical performance metrics and towards verifying agents' present capabilities and potential to deliver desired outcomes, facilitating a comprehensive verification of trust.

The exploration of trust verification techniques has higher-level consequences that extend beyond individual assessments of trust in MAS. While reputation-based trust models provide valuable insights, it is important to acknowledge that relying solely on reputation may result in biased misplaced trust. By delving into the intricacies of trust and developing reliable verification methods that look at temporal dynamics of trust and allow reasoning about what agents can deliver prospectively, we can pave the way for ethical AI systems. These systems allow users to foresee what other (AI) agents can de-

[3]Note that relying on the past may also support more bias in reasoning about other agents in human-AI systems.

liver, compare those against their own preferences and values, and actively reason about and verify trust in a reliable and transparent manner. To address these critical aspects, we require temporally expressive tools to represent and reason about trust dynamics. In the next section, we survey such formal tools and highlight weaknesses and points we learned for formulating our perspective on actual trust in MAS.

3. Formal Methods for Temporal Trust Reasoning

We re-iterate that actual trust is forward-looking and requires formal methods for temporal reasoning about future computations occurring in a MAS. We therefore direct our focus here on works where temporality is a key aspect of reasoning about trust. To this end, little is explored in the literature for performant techniques for verifying temporal trust properties in MAS. Drawel et al. [4] address the problem of model checking trust logics, and propose a model checking technique for two temporal logics for trust, namely TCTL, a (pre-conditional) Computation Tree Logic of Trust and TCTL^c for conditional trust, which extend CTL to introduce new modalities to reason about trust. They ground the semantics of their trust logics on a formalism of interpreted systems [19] which is enriched with a trust function – a binary relation between two states, which associates to each local state of each agent the "trust vision" the trustor has towards other agents in a given global state. States are compatible with each other with regard to the trust an agent has with another. The (pre-conditional) trust modality $T_p(i, j, \psi, \varphi)$ in TCTL stands for "Preconditional Trust" and is read as "the trustor i trusts the trustee j to bring about φ given that the precondition ψ holds"; ψ holds in a state s, and φ needs to hold in all trust-accessible states between agents i and j different from s. On the other hand, the conditional trust modality $T_c(i, j, \psi, \varphi)$ in TCTL^c is read as "agent i trusts agent j about the consequent φ when the antecedent ψ holds". Differently from pre-conditional trust where the precondition ψ must hold before the trust content φ is brought about, conditional trust needs at least one trust-accessible state satisfying ψ. Agents are specified in their model using the VISPL [6] language which extends ISPL [20] used for specifying traditional interpreted systems. A transformation-based procedure is used to convert the model checking problem of TCTL and TCTL^c into model checking CTL. New symbolic model checking algorithms are proposed to formally specify and automatically verify a system under consideration against properties in TCTL and TCTL^c. The semantics of their trust logics are interpreted in a vector-extended version of interpreted systems, capturing the trust relationships between interacting parties within the model itself. We argue that assuming trust relations as known in advance is not realistic and is a phenomenon to be reasoned about automatically as we do here. For model checking, transformations to CTL-supported models and formulae are exploited to make use of the NuSMV model checker CTL [21].

Inspired by Falcone and Castelfranchi's [14] cognitive notion of social trust, Huang and Kwiatkowska [13] introduce a framework for quantifying and reasoning about cognitive trust, governing social relationships between humans and autonomous systems. A semantic model grounded on stochastic games is introduced, namely Autonomous Stochastic Multi-Agent Systems (ASMAS). Differently to [4] where trust is represented as a binary relation, probabilistically quantified degrees of trust are expressed in terms of belief. The ASMAS model combines a stochastic game, where agents only have partial

observability of the state, with a mechanism providing agents with goals, intentions and preferences. An ASMAS differentiates between transitions in the temporal and cognitive space; the actions of the agent occur in physical space, whereas cognitive processes lead to changes in the agent's mental state, and lead to decisions about which (physical) action to take. Properties are specified using PRTL* (Probabilistic Rational Temporal Logic), an extension of the probabilistic temporal logic PCTL* which introduces cognitive attitude and trust operators. In PRTL*, cognitive operators $\mathbb{G}_A\psi$ (ψ holds in the future regardless of agent A changing its goals), $\mathbb{I}_A\psi$ (ψ holds in the future regardless of A changing its intentions), $\mathbb{C}_A\psi$ (agent A can change its intention to achieve ψ), are used with respect to an agent A and a task ψ; they quantify over the possible changes of goals, intentions and available intentions respectively. For $\bowtie \in \{<,\leq,>,\geq\}$ and $q \in \{0,1\}$, the belief operator $\mathbb{B}_A^{\bowtie q}\psi$ (agent A believes ψ with probability in relation $\bowtie$ to q) probabilistically quantifies agent A's belief of ψ, the competence trust operator $\mathbb{CT}_{A,B}^{\bowtie q}\psi$ (agent A trusts agent B with probability in relation $\bowtie$ with q on its *capability* of completing task ψ) probabilistically quantifies the degree of agent A's trust in B's capability to achieve ψ and that there exists a valid intention for B to carry out φ, and the disposition trust operator $\mathbb{DT}_{A,B}^{\bowtie q}\psi$ expresses that agent A trusts agent B with probability in relation $\bowtie$ with q on its *willingness* of completing the task ψ; that is, ψ is unavoidable for all intentions. These operators probabilistically quantify an agent's beliefs, the degree of agent A's trust in agent B's capability and willingness to carry out ψ respectively for each case. The verification problem for full PRTL* is found to be undecidable, but with decidable fragments.

As seen from the above works, verifiable notions of actual trust need to be forward-looking and require practical formal methods for temporal reasoning. Our learnings point us towards our formulation on actual trust in MAS.

4. Modelling Interactions in Multiagent Systems

In this section we discuss the basic building blocks necessary to reason about trust in a MAS. An *interpreted system* (IS) [19] is a model for representing a MAS; it is a formal description of the computations carried out by a set of agents.

Definition 1 (Interpreted Systems)**.** *For a set of agents* $Agt = \{1,\ldots,n\}$*, where* n *is the number of agents in the system, an interpreted system IS is a tuple:* $IS = \langle (L_i, Act_i, P_i, t_i)_{i\in Agt\cup\{E\}}, I, h\rangle$*, where:*

- *Each agent* $i \in Agt \cup \{E\}$ *is characterised by a finite set of private local states* L_i*, which determines all information relevant to agent* i *in a given global state.*
- Act_i *is a finite set of actions that may be performed for agent* i*.*
- $P_i : L_i \to 2^{Act_i \setminus \varnothing}$ *is a protocol for agent* i*. The actions of the agent must be performed in compliance with the protocol, which allows for non-determinism in the system when more than one action is enabled for a given local state.*
- E *is a special "agent", referred to as the environment. It has its own sets of local states* L_E*, actions* Act_E*, protocol* P_E *and transition function* t_E*.*

- $t_i : L_i \times Act_1 \times \cdots \times Act_n \times Act_E \to L_i$ *is a (partial) transition function describing the (deterministic) evolution of agent i's local states.* [4] *Every action is assumed to be protocol-compliant; if* $l_i' = t_i(l_i, a_1, \ldots, a_i, \ldots, a_n, a_E)$*, then* $a_i \in P_i(l_i)$ *for all* $i \in Agt \cup \{E\}$.
- I *is a set of initial global states.*
- $h : AP \to 2^Q$ *is a valuation function defining the set of states where certain atomic propositions are true, where AP is a set of atomic propositions and* $Q = L_1 \times \cdots \times L_n \times L_E$ *is a set of global states.*

Local states cannot be observed by other agents. Actions are observable by other agents i.e. to determine an agent's transition function t_i. *Global states* Q combine the local states of all $n+1$ agents. *Global actions* $ACT = Act_1 \times \cdots \times Act_n \times Act_E$ combine all agents' action sets. Consider an agent $i \in Agt \cup \{E\}$. For a global state $q = (l_1, \ldots, l_n, l_E) \in Q$ the function $loc_i : Q \to L_i$ where $loc_i(q) = l_i$ returns the local state of agent i in global state q, and for a joint action $a = (a_1, \ldots, a_n, a_E) \in ACT$ the function $act_i : ACT \to Act_i$ is such that $act_i(a) = a_i$ returns the action of agent i in joint action a. A *global transition function* defines the means through which the system evolves through time.

Definition 2 (Global Transition)**.** *Given two global states* $g, g' \in Q$ *and a joint action* $a \in ACT$*, a (partial)* global transition *function* t *is such that* $t(g,a) = g'$ *iff* $t_i(loc_i(g), act_i(a)) = loc_i(g')$ *and* $act_i(a) \in P_i(loc_i(g))$ *for all* $i \in Agt \cup \{E\}$*. This property can be abbreviated* $g \to_a g'$ *iff* $t(g,a) = g'$.

Since each t_i is deterministic, at most one such transition exists for each g and a. A joint action $a \in ACT$ is enabled in state $q \in Q$ if there exists a state $q' \in Q$ such that a transition exists between q and q' through the execution of a. A Kripke model $\mathcal{M}_{IS}$ associated to an interpreted system IS defines its semantics when interpreting temporal formulae. Knowledge of agent i is represented in terms of an indistinguishability relation $\sim_i$, an equivalence relation with known properties: reflexivity, symmetry and transitivity. Informally, a property φ being "known" by an agent i is determined by φ holding in all global states indistinguishable by i.

Definition 3 (Associated Kripke model)**.** *A Kripke model* $\mathcal{M}_{IS}$ *associated with* $IS = \langle (L_i, Act_i, P_i, t_i)_{i \in Agt \cup \{E\}}, I, h \rangle$ *is a tuple* $\mathcal{M}_{IS} = (W, R, h, \{\sim_i\}_{i \in Agt})$ *such that worlds* $w \in W$ *are the global states of IS reachable via the set of initial states I through the transition relation R. Two worlds* w, w' *are related by the transition relation* $R \subseteq W \times W$ *when there is a joint action a such that* $w \to_a w'$*. The valuation function h is used as a* labelling function*, and the* epistemic indistinguishability relation *[19, p. 117] for agent* $i \in Agt$*, namely* $\sim_i \subseteq W \times W$*, relates a pair of global states* $w, w' \in W$ *whenever agent i has the same local state in both w and* w'*. In other words, it cannot distinguish w and* w'.

A (potentially infinite) *path* π is defined as a sequence of states where each pair of successor states are related by a global transition; that is, a path is a sequence of states $\pi = (q^0, q^1, \ldots, q^n, \ldots)$ such that for all $i \geq 0$, we have $(q^i, a^i, q^{i+1}) \in t$ for some action $a^i, \ldots \in ACT$. The state at position k is denoted $\pi(k)$. Q_Γ denotes the projection of Q on the local states of the agents in $\Gamma \subseteq Agt \cup \{E\}$ and similarly ACT_Γ for the elements of ACT restricted to the agents in Γ. For example, if $a = (a_1, a_2, a_3, a_4, a_5, a_E) \in ACT$ and we

[4] Here we adopt the definition with $n+1$ evolution functions as used in [22]. Each evolution function gives the next local state as a function of the current local state of the agent and all the other agents' actions.

take $\Gamma = \{1,3,5\}$, then an element a_Γ of ACT_Γ is $a_\Gamma = (a_1, a_3, a_5)$. A *strategy* provides the semantics in an IS of *strategic operators* in temporal formulae such as those encountered in ATL [23].

Definition 4 (Strategy)**.** *For an agent i of an IS, a (memoryless) strategy s_i is a function $s_i : L_i \to 2^{ACT_i} \setminus \{\varnothing\}$ such that if $a_i \in s_i(l_i)$, then $a_i \in P_i(l_i)$.*

A strategy depending on a history, or sequence, of local states is known as a *memory-based* strategy. In addition to this, note that agents may perform different actions in different global states whose local component is the same, allowing for non-deterministic strategies. This is known as a *non-uniform* strategy. As is done in [20], we also focus on non-uniform, memoryless, incomplete information strategies here. In a MAS, *joint strategies* are a collection of individual strategies[5].

Definition 5 (Joint Strategy)**.** *Given a coalition Γ, a* joint strategy *for Γ is a function $s_\Gamma : Q_\Gamma \to 2^{ACT_\Gamma} \setminus \{\varnothing\}$ such that $s_\Gamma(l_{x_1}, \ldots, l_{x_k}) = (s_{x_1}(l_{x_1}), \ldots, s_{x_k}(l_{x_k}))$, where $s_{x_1}, \ldots, s_{x_k}$ are strategies for the agents $x_1, \ldots, x_k \in \Gamma$.*

Intentions in MAS. We consider *intentions* as goals that each agent intends to deliver. They are declared publicly in terms of statements in propositional logic.

Definition 6 (Interpreted systems with intentions)**.** *We define an interpreted system with intentions as a tuple $ISI = \langle (L_i, Act_i, P_i, t_i, \mathcal{I}_i)_{i \in Agt \cup \{E\}}, I, h \rangle$, where a consistent set of* intentions *$\mathcal{I}_i \subseteq 2^\Phi \setminus \varnothing$ are such that each agent i is associated to a finite set of k propositions $\{\varphi_1, \ldots, \varphi_k\}$, with each $\varphi \in \Phi$ being propositional formulae, that it 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 the set of intentions for individual agents intentions; $p \in \mathcal{I}_i \Rightarrow \neg p \notin \mathcal{I}_i$ for all $i \in Agt$, for any proposition p. It is not possible for an agent to have the intention to go out and stay at home simultaneously. The model $\mathcal{M}_{ISI}$ associated with the interpreted system with intentions *ISI* is defined identically to that in Definition 3. Note that an interpreted system is a special case of an ISI, where $\mathcal{I}_i = \top$ for all $i \in Agt$. Unlike in the treatment introduced in [24], intentions 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.

Alternating-time temporal logic. Alternating-time temporal logic (ATL) [23] generalises CTL and is used for strategic reasoning in MAS. It is used to describe what a collection of agents can achieve. Similarly to CTL, we have the usual atomic propositions, negation, binary conjunction and disjunction operators p, $\neg$, $\wedge$ and $\vee$ respectively. In ATL, given a set of agents Agt, a *coalition* $\Gamma \subseteq Agt$ and a property φ, the specification $\langle\!\langle \Gamma \rangle\!\rangle X\varphi$ is read as: "the coalition Γ have a joint strategy to achieve φ in the next step independently of what $Agt \setminus \Gamma$ does."

Here and in [20] we interpret the semantics of ATL on the temporal model $\mathcal{M}_{IS}$, given an initial state $q^0 \in Q$, a formula φ and a set of atomic propositions AP where $p \in AP$. The semantics for non-temporal operators are equivalent to that for CTL. We

[5] When in the context of speaking about a collection of agents, we refer to joint strategies simply as strategies.

here will focus on the semantics of the "next" operator X only: $(\mathcal{M}_{IS}, q^0) \vDash \langle\langle\Gamma\rangle\rangle X\varphi$ iff there exists a joint strategy s_Γ and joint action $a_\Gamma \in s_\Gamma(q^0_\Gamma)$ such that for all actions a whose restriction to Γ is equal to a_Γ and for all states q^1 such that $q^0 \to_a q^1$ we have that $(\mathcal{M}_{IS}, q^1) \vDash \varphi$. Connectives for reasoning about strategic ability over sequences of states are not discussed here.

ATLK [20] combines ATL with modal operators to reason about the knowledge of agents in a MAS. Here we focus on a fragment which is critical to define trust modalities consisting of those given in "Vanilla ATL" with the knowledge operator $K_i\varphi$, which is read as "agent i knows φ". More precisely, for a model $\mathcal{M}$, state q^0 and property φ, $(\mathcal{M}, q^0) \vDash K_i\varphi$ iff for all $q^1 \in Q$ we have that $q^0 \sim_i q^1$ implies that $(\mathcal{M}, q^1) \vDash \varphi$. In other words, for agent i there isn't a state indistinguishable from q^0 where φ does not hold. The agent has enough information in its own local states to determine from its perspective that φ holds in the system.

We note that 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. Although we do not focus on complexity in this work, the semantics of ATLK in this context are analogous to the imperfect information, memoryless strategies case of ATL, namely ATL$_{ir}$ [25,26]. The model checking problem for ATL$_{ir}$ is decidable and Δ_2^P-complete.

5. A Computational Notion of Actual Trust

The semantics defined in the previous section form the basis of that which will be used to reason about actual trust, where an agent trusts other agents to collectively perform a task, which we encode here in terms of ATLK formulae. This will aid in being able to transform from the trust verification problem into an ATLK model-checking problem. As common in formal methods research, we evaluate our notion's properties formally and show applicability in a well-established running example to showcase the expressiveness of our formal notion of actual trust for reasoning about different aspects of trust in multiagent systems. We call the specification language $\mathcal{L}$. Assume that $\mathcal{L}$ contains the standard Boolean connectives of CTL. In terms of the trustee β (a group of potentially trusted agents), the trustor α and task T (see Section 1), we take an agent i as α, the group of agents Γ as β 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". Specifically:

Definition 7 (I trust Γ if I know they can deliver)**.** *Given a model associated with an interpreted system with intentions $\mathcal{M}_{ISI}$ and an agent $i \in Agt$, we say that $(\mathcal{M}_{ISI}, q^0) \vDash \mathcal{T}_i(\Gamma, \varphi)$ 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 $q^K \to_a q^1$, we have that $\varphi \cap \bigcap_{i\in\Gamma} \mathcal{I}_i$ is nonempty and consistent, and $(\mathcal{M}_{ISI}, q^1) \vDash \varphi$.*

That is, φ is consistent with each agent's intentions. We note that with this definition, trust 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 intentions that all agents intend to deliver.

It is permitted for $i \in \Gamma$ or $\Gamma = \{i\}$, where agent i trusts that it can cooperate with the agents in Γ to bring about φ, and that agent i has trust in itself that it can bring about φ respectively, regardless of what the agents in $Agt \setminus \Gamma$ do. From this, it is possible to reason about supersets of agents:

Proposition 1 (Non-monotonicity of trust)**.** *Let* $\Gamma \subseteq \Gamma'$. *Then* $\mathcal{T}_i(\Gamma, \varphi) \not\Rightarrow \mathcal{T}_i(\Gamma', \varphi)$.

Proof. Assume a $\mathcal{M}_{ISI}$ with $Agt = \{1,2,3\}$, $\Gamma = \{1,2\}$, and intentions $\mathcal{I}_1 = \{\varphi, \psi\}$, $\mathcal{I}_2 = \{\varphi\}$ and $\mathcal{I}_3 = \{\neg\varphi\}$. Each $\mathcal{I}_i$ is clearly consistent. Now, without loss of generality, let $i = 1$ and assume $q \vDash \mathcal{T}_1(\Gamma, \varphi)$ for some state $q \in Q$. Assume $\Gamma' = \Gamma \cup \{3\} = \{1,2,3\}$. For $q \vDash \mathcal{T}_1(\Gamma', \varphi)$ to hold, it would mean that $\varphi \cap \bigcap_{i \in \Gamma} \mathcal{I}_i = \varphi \cap \{\varphi, \psi\} \cap \{\varphi\} \cap \{\neg\varphi\} = \varnothing$ is consistent, which is a contradiction. □

This highlights the importance of considering intentions – without them, a notion of trust could show unintuitive results, but with the inclusion of intentions, not all supersets of a set of agents can be trusted for φ. We illustrate our notion of trust in The Bit Transmission Problem (BTP) [19, p. 114].

Example 1 (Bit Transmission Problem)**.** *In the bit transmission problem, a sender S wants to communicate the value of a bit to a receiver R over a faulty communication channel. Messages between S and R may be lost, but the value of the bit will not be corrupted. One needs to define a protocol for S to be sure that R received the bit. An example protocol is the following: S sends the value of the bit to R and continues to do so until it receives an acknowledgement ("ack"), after which it will stop sending the value of the bit. R does nothing until it receives the value of the bit, and then it sends acknowledgements to S forever in the future.*

We encode the BTP as an ISI in a standard way; by considering the local states of the sender and receiver agents S and R, and the environment agent E, which will be used to represent the faulty communication channel. The sender's state will consist solely of the value of the bit, or the value of the bit combined with the acknowledgement sent from the receiver. The local states representing the situations where the sender has received the acknowledgement are denoted $0{-}ack$ and $1{-}ack$. The receiver was either sent the value of the bit, with its local state equal to the bit's value, or it is empty, denoted by ε. The environment state does not play a role in our formalisation of the bit transmission problem, so we take $L_E = \{\cdot\}$. We now have the following local states: $L_S = \{0, 1, 0{-}ack, 1{-}ack\}$, $L_R = \{0, 1, \varepsilon\}$ and $L_E = \{\cdot\}$. Omitting the environment's local state, this generates six global states: $(0,\varepsilon)$, $(0,0)$, $(0{-}ack, 0)$, $(1,\varepsilon)$, $(1,1)$, and $(1{-}ack, 1)$. Consider the propositional atom **recack**, representing all global states where the receiver was successfully transmitted the bit's value and the sender has received the acknowledgement, such that $h(\mathbf{recack}) = \{(0{-}ack, 0), (1{-}ack, 1)\}$. Then, the interpreted system IS consisting of the agents S, R and E satisfies **recack** at any global state $q \in I$, where I is the set of initial states containing the two global states with the sender's local component of q being either $1{-}ack$ or $0{-}ack$.

Assume in the corresponding ISI that all agents intend for acknowledgements to always be received, i.e. $\mathcal{I}_S = \mathcal{I}_R = \mathcal{I}_E = \mathbf{recack}$. It is easy to check whether $IS, q \vDash T_S(R, \mathbf{recack})$, i.e. the sender trusts the receiver in bringing about **recack**. Intuitively, starting from the initial state q of either $(0,\varepsilon)$ or $(1,\varepsilon)$, the only states accessible via the sender's accessibility relation $\sim_S$ are $(0,0)$ and $(1,1)$ respectively. For both of these

states, there exists a strategy for R such that regardless of what S does, states satisfying **recack** follow, since the receiver can always send the acknowledgement to the sender regardless of the value of the bit in the previous state. If the faulty communication channel is modelled in the environment, and in every state it prevents the receiver from sending the acknowledgement, the formula would not hold, as a suitable strategy for R does not exist, even though the environment intended that acknowledgements should always be received since $\mathcal{I}_E$ = **recack** in the model.

Expanding upon this notion of trust, our modelling also facilitates reasoning about trusting agents who employ multistep strategies to ensure the realisation of a state of affairs φ. While our formalisation emphasises the immediate outcomes, our model acknowledges the potential for agents with longer-term strategies to be trusted as well. By verifying the effectiveness of multistep strategies, agents can evaluate and place trust in individuals or groups who demonstrate the ability to "*eventually*" achieve a desired outcome φ. This extended perspective on trust enables a more comprehensive analysis of trust dynamics in complex scenarios, accommodating both immediate and long-term strategies for attaining desired goals as well as providing a base for quantifying trust (e.g., agent i may trust a Γ who can ensure φ in the immediate next state more than Γ' with a multistep strategy to do so).

6. 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 the fulfilment of a task. We capture the epistemic dynamics of trust and applicability for reasoning about trust in real-world scenarios.[6] We use the standard knowledge and strategic operators $K_a\varphi$ and $\langle\langle a \rangle\rangle\varphi$ assuming the "de dicto" semantics of knowledge: an agent only knows that a strategy is *available*. Outside the scope of this work is to consider a stronger view of strategic ability, e.g. such as that introduced in [27] where an agent also knows the specific strategy.

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. The framework

[6] We highlight that as we modelled our notions in ATL, verifying actual trust can be implemented in standard model-checking tools such as MCMAS [20].

of interpreted systems also allow for a group of external observers to be modelled either as the environment "agent" E in a similar sense to [28, p. 10] using the trivial protocol function (returning a no-op action for all local states). By recognising such relationships between individual and collective trust, we gain a better understanding of trust dynamics in human-AI systems. We can analyse how individual-level trust influences coordination within groups. This understanding is crucial in various contexts, such as teamwork, organisational dynamics, and social networks, where trust plays a pivotal role in achieving common goals and fostering collaboration.

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. This allows for the exploration of various trust dynamics and the potential for trust to evolve over time, reflecting the nuanced nature of human-AI interactions and decision-making.

This framework is also compatible with systems with machine learning components. In particular, neural interpreted systems [29] with neural networks used for perception tasks combined with a symbolic controller unit. This can allow us to use traditional verification techniques while exploiting the recent advances in neural network analysis techniques such as [30,31].

7. Concluding Remarks and Future Contributions

In this paper we demonstrate the need to establish the notion of trust in multiagent systems consisting of both human and AI agents. We outlined a method to capture trust using alternating-temporal logic with knowledge and intentions, and exemplified the approach in a toy example. Similarly to recent approaches to the trust verification problem, our verifiable notion of trust allows for a transformation into a related tractable model checking problem to be analysed by existing standard model checking tools such as MCMAS, which we aim to use to empirically evaluate the approach for future work. In addition, we argue that verifying actual trust is less biased than performing an analysis of the reputation of agents and their past behaviour. We wish to explore different notions of trust, supporting multistep strategies, and eventually curate a framework for reasoning about trust, allowing also for quantification [32]. We will also utilise Event-B [33,34,35] to explore *refinement-based* [36] formal methods for actual trust.

Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) through the UKRI Trustworthy Autonomous Systems Hub (EP/V00784X/1) and a Turing AI Fellowship (EP/V022067/1) on Citizen-Centric AI Systems. We would also like to thank the anonymous reviewers for their helpful comments, which have substantially improved the paper.

References

[1] Ramchurn SD, Stein S, Jennings NR. Trustworthy human-AI partnerships. Iscience. 2021;24(8):102891.

[2] Akintunde M, Yazdanpanah V, Fathabadi AS, Cirstea C, Dastani M, Moreau L. Actual Trust in Multiagent Systems. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024). International Foundation for Autonomous Agents and Multiagent Systems; 2024. p. 1-3.

[3] Halpern JY. Actual causality. Cambridge, Massachusetts, United States: MIT Press; 2016.

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

[5] Drawel N, Bentahar J, Laarej A, Rjoub G. Formal verification of group and propagated trust in multi-agent systems. Autonomous Agents and Multi-Agent Systems. 2022;36(1):19.

[6] Drawel N, Bentahar J, Shakshuki E. Reasoning about Trust and Time in a System of Agents. Procedia Computer Science. 2017 12;109:632-9.

[7] Bentahar J, Drawel N, Sadiki A. Quantitative Group Trust: A Two-Stage Verification Approach. In: Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems (AAMAS'22). Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems; 2022. p. 100-8.

[8] Cohen PR, Levesque HJ. Intention is Choice with Commitment. Artif Intell. 1990;42(2-3):213-61. Available from: https://doi.org/10.1016/0004-3702(90)90055-5.

[9] Burnett C, Norman TJ, Sycara K. Trust decision-making in multi-agent systems. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI'11). Barcelona, Catalonia, Spain: AAAI Press; 2011. p. 115-20.

[10] Ramchurn SD, Huynh D, Jennings NR. Trust in multi-agent systems. The knowledge engineering review. 2004;19(1):1-25.

[11] Gladyshev M, Alechina N, Dastani M, Doder D. Group Responsibility for Exceeding Risk Threshold. In: Proceedings of the 20th International Conference on Principles of Knowledge Representation and Reasoning, KR 2023, Rhodes, Greece, September 2-8, 2023; 2023. p. 322-32. Available from: https://doi.org/10.24963/kr.2023/32.

[12] Herzig A, Lorini E. A Dynamic Logic of Agency I: STIT, Capabilities and Powers. Journal of Logic, Language, and Information. 2010;19(1):89-121.

[13] Huang X, Kwiatkowska M. Reasoning about Cognitive Trust in Stochastic Multiagent Systems. In: Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17). San Francisco, California, USA: AAAI Press; 2017. p. 3768-74.

[14] Falcone R, Castelfranchi C. Social Trust: A Cognitive Approach. In: In Trust and Deception in Virtual Societies. Berlin: Springer; 2001. p. 55-90.

[15] Ramchurn SD, Jennings NR, Sierra C, Godo L. Devising a trust model for multi-agent interactions using confidence and reputation. Applied Artificial Intelligence. 2004;18(9-10):833-52.

[16] Ågotnes T, Goranko V, Jamroga W, Wooldridge M. Knowledge and ability. Handbook of Epistemic Logic. 2015.

[17] Mousavi MR, Cavalcanti A, Fisher M, Dennis L, Hierons R, Kaddouh B, et al. Trustworthy Autonomous Systems Through Verifiability. Computer. 2023;56(2):40-7.

[18] Robinette P, Li W, Allen R, Howard AM, Wagner AR. Overtrust of robots in emergency evacuation scenarios. In: Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16). Christchurch, New Zealand: IEEE; 2016. p. 101-8.

[19] Fagin R, Halpern JY, Moses Y, Vardi MY. Reasoning about Knowledge. Cambridge: MIT Press; 1995.

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

[21] Cimatti A, Clarke E, Giunchiglia F, Roveri M. NuSMV: A New Symbolic Model Verifier. In: Proceedings of the 11th International Computer Aided Verification Conference. Trento, Italy: Springer; 1999. p. 495-9.

[22] Raimondi F. Model checking multi-agent systems [Ph.D. thesis]; 2006. Available from: http://discovery.ucl.ac.uk/5627/.

[23] Alur R, Henzinger TA, Kupferman O. Alternating-Time Temporal Logic. Journal of the ACM. 2002;49(5):672-713.

[24] Jamroga W, van der Hoek W, Wooldridge M. Intentions and strategies in game-like scenarios. In: Progress in Artificial Intelligence: 12th Portuguese Conference on Artificial Intelligence, EPIA'05. Cov-

ilhã, Portugal: Springer; 2005. p. 512-23.
[25] Schobbens PY. Alternating-time logic with imperfect recall. Electronic Notes in Theoretical Computer Science. 2004;85(2):82-93.
[26] Jamroga W, Dix J. Model Checking Abilities of Agents: A Closer Look. Theory of Computing Systems. 2008;42(3):366-410.
[27] Jamroga W, Ågotnes T. Constructive knowledge: what agents can achieve under imperfect information. Journal of Applied Non-Classical Logics. 2007;17(4):423-75.
[28] Raimondi F, Lomuscio A. A tool for specification and verification of epistemic properties in interpreted systems. Electronic Notes in Theoretical Computer Science. 2004;85(2):176-91.
[29] Akintunde M, Botoeva E, Kouvaros P, Lomuscio A. Verifying Strategic Abilities of Neural-symbolic Multi-agent Systems. In: Proceedings of the 17th International Conference on Principles of Knowledge Representation and Reasoning (KR20). IJCAI Press; 2020. p. 22-32.
[30] Katz G, Huang DA, Ibeling D, Julian K, Lazarus C, Lim R, et al. The Marabou Framework for Verification and Analysis of Deep Neural Networks. In: Proceedings of the 31st International Conference on Computer Aided Verification (CAV19); 2019. p. 443-52.
[31] Singh G, Gehr T, Püschel M, Vechev P. An abstract domain for certifying neural networks. Proceedings of the ACM on Programming Languages. 2019;3(POPL):41.
[32] Yazdanpanah V, Dastani M. Quantified degrees of group responsibility. In: Coordination, Organizations, Institutions, and Norms in Agent Systems. Cham: Springer; 2016. p. 418-36.
[33] Abrial JR. Modeling in Event-B: System and Software Engineering. Cambridge, UK: Cambridge University Press; 2010.
[34] Lanoix A. 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). Nanjing, China: IEEE Computer Society; 2008. p. 297-304.
[35] Gao HJ, Qin Z, Lu L, Shao LP, Heng XC. Formal specification and proof of multi-agent applications using event b. Information Technology Journal. 2007;6(7):1181-9.
[36] Fathabadi AS, Yazdanpanah V. Trust modelling and verification using Event-B. In: Proceedings of the Fifth Workshop on Formal Methods for Autonomous Systems (FMAS'23). Leiden, Netherlands: EPTCS; 2023. p. 10-6.