A Computational Model of Trustworthiness: Trust-Based Interactions Between Agents in Multi Agent System

Basten Leeftink, Britta Abbink Spaink, Tomasz Zurek[a] and Tom van Engers[b]

Complex Cyber Infrastructure, Informatics Institute, Faculty of Science, University of Amsterdam, Netherlands
{britta.abbink.spaink, basten.leeftink}@student.uva.nl, {t.a.zurek, T.M.vanEngers}@uva.nl

Keywords: Trust, Trustworthiness, Agent-Based Programming.

Abstract: In our research group working on normative systems, we develop (Normative) Agent Based Models for evaluating policies, and as a basis for building distributed (normative) control components. If and how interactions between actors (represented by agents) take place are heavily impacted by the (dis)trust between those actors. In this paper, we discuss a model of the representation of the three components of the agent's trustworthiness: competence, benevolence, and integrity. The model presented in this paper is being illustrated by a small simulation experiment.

1 INTRODUCTION

Trust between actors in social systems is an essential factor that influences if and how the actors interact. In socio-technical systems, where the behaviour of devices or components should act in the interests of the stakeholders involved, computational trust models serve as a mechanism to establish relationships between these devices/components or not, hence impacting the (emerging) functionalities of such socio-technical systems.

In this paper we describe how we combine models of trust with agent-based modelling. While, as stated before, we aim to use our ABMs as a basis for creating control components, we also have a more theoretical purpose in mind. As trust is an important if not determinant factor in social and social-technical systems, our computational modelling approach allows for (dis-) confirmation of the existing social trust theories, by modelling these theories in a Multi-Agent Systems (MASs) simulation (using ABM) and compare the results with actually observed phenomena. We focus on testing whether it is possible to model trust phenomena in a MAS in a way that reflects the trust-related phenomena observable in society. In silicon experiments (i.e. simulations) can only lead to accurate results if all relevant trust factors are adequately modelled in the ABM used for that simulation.

[a] https://orcid.org/0000-0002-9129-3157
[b] https://orcid.org/0000-0003-3699-8303

In this paper we will focus on the problem of the representation of trustworthiness of potential trustees. Trustworthiness is one of the basic components of trust, which is the basis of the Trustor's decision concerning placing trust in a particular agent. This concept is an object of an extensive research in social sciences (Castelfranchi and Falcone, 2010; Mayer et al., 1995; Mcknight et al., 2011), Multi-agent Systems (Delijoo, 2021), and others. In our work, we explore this concept in order to prepare a comprehensive model and implementation in MAS.

In this paper we introduce a computational model of trustworthiness of one agent (Trustee) in the eyes of other agent (Trustor). We will address both the ontological aspects as well as the epistemological aspects of this model. Other aspects such as the mechanisms of building and eliminating trust between agents, the role of evidence, the creation of plans of agents, delegating tasks, etc. are left out of scope, and will be addressed in future work.

2 TRUST MODELS IN LITERATURE

Before going into some influential trust models found in literature we should point at different approaches for creating MASs. MASs are usually constructed on the basis of one of the general AI paradigms: either they are knowledge-driven (typically created with the use of Agent-based programming), or data driven

(typically constructed on the basis of multi agent reinforcement learning mechanisms). Below we shortly present a discussion of various approaches.

Nobandegani et al. (Nobandegani et al., 2023) present a model of trust for Multi-agent reinforcement learning. Their paper presents a rather oversimplified reputation-based trust mechanism in which trust is built on the basis of past experience, without differentiating what type of situations the agents face, what types of tasks are relevant for agents, without individual attitudes of agents, etc. In their model, trust is represented as a number. Also Tykhonov et al. represent trust as a number (Tykhonov et al., 2008). In their paper they present an agent-based model of trust, created in order to analyse trust dynamics in a specific supply-chain experiment. Chen on the other hand, (Chen et al., 2015) represents trust binary (trustor trusts trustee or not). In their paper they describe some experiments on modeling trust games with the use of agent-based modeling. Similar to Nobandegani, trust is build on the basis of past experience, but their model distinguishes so-called myopic trust in which experience is limited to one step back (short-term memory). Jaffry and Treur (Jaffry and Treur, 2013) model trust with the use of agent-based modeling, again trust is represented as a number calculated on the basis of experience. Their model does not take into consideration all complex aspects of trust, like trustworthiness, the role of competence, benevolence, and integrity in the trust evaluation mechanism. Parsons, et al. (Parsons et al., 2012) introduce a model of reasoning about trust in an argumentation framework, but trust in this model is still a single number only. The model introduces an interesting view on the propagation of trust in the argumentation, but the authors avoid the discussion about the nature of trust, its individualised and societal character, etc.

Frey and Martinez (Frey and Martinez, 2024) notice some important drawbacks of existing approaches in which trust is reduced to reputation. Following that observation they introduce a new way of representing trust in a multi-agent reinforcement learning-based system. Although their approach extends the reputation-based models, it is based on a particular and quite specific understanding of trust which reduces its role to learning to follow someone's else (trustee) behaviour, rather than to represent the attitude of a trustor towards trustee. Although the model is claimed to simulate some important elements of mammal's neural mechanisms (the dopamine production, higher order conditioning), it ignores other trust-related phenomena, particularly at the social-interaction level, like task dependency, the evaluation of benevolence and integrity of trustee, etc. Fung et

al. also (Fung et al., 2022) present a model of trust for reinforcement learning-based multi agent system. Their model aims at representing the mechanism of finding consensus between agents. Their trust model is also quite oversimplified: there is a binary relation of trust between agents (trust/ not trust), and it lacks any in depth analysis of the nature of the phenomenon trust.

3 THE COMPLEX NATURE OF TRUSTWORTHINESS

Studying the most cited definitions and theories concerning the trust relations, learn us that despite the tendency to express Trust as a number or a boolean, Trust and Trustworthiness are in fact complex concepts, based upon some primitives. In this section we will discuss some of the primitives (concepts/dimensions) suggested in literature.

Let's start with the definition that can be found in the Merriam-Webster Dictionary that defines trust as: *assured reliance on the character, ability, strength, or truth of someone or something*, but there are a number of other definitions.

Castelfranchi and Falcone (Castelfranchi and Falcone, 2010) strictly distinguish between trust and trustworthiness: trust is a property of the trustor (towards a trustee), while trustworthiness is a property of the trustee. The stance of Castelfranchi and Falcone however, is not unproblematic, as while being a property of the trustee that property is attributed by the trustor, making it a subjective concept. Trustworthiness could be misplaced by a wrong evaluation (belief) of trustworthiness in the trustor's mind. Trustworthiness in their model remains to be a subjective as well as a relational concept, i.e. the trustworthiness of a person does not only depend on that person but also on the person they have a relationship with. According to Castelfranchi and Falcone, Trust is built on the basis of Trustworthiness, but it also depends on the personality of the Trustor, the plausibility gap (presence or lack of evidence), etc. Some other authors (see e.g. (Mayer et al., 1995)) point out some key components of trustworthiness: competence, benevolence, and integrity. Others, like McKnight et al. (Mcknight et al., 2011) distinguish competence, benevolence, and predictability (of behaviour). Following that, (Castelfranchi and Falcone, 2010) decompose competence into two dimensions: the evaluation of *skills* and *know-how* (knowledge of recipes, techniques, etc), while the concept of willingness (which can be interpreted as benevolence) is represented by *concern* or *certainty of adoption* and

persistence.

Meyer et. al. (Mayer et al., 1995) introduce a specific architecture of trust in which factors of perceived trustworthiness (competence, benevolence, and integrity) along with the trustor's propensity build trust. Delijoo (Delijoo, 2021) extended this model by adding the influence of context both on the evaluation of trustworthiness and the general decision of trust. In this paper, we have extended this model and incorporate it into a more general architecture of agents. Castelfranchi and Falcone (Castelfranchi and Falcone, 2010) distinguish three 'levels of trust': trust as mental attitude, trust as decision, and trust as action. Although we agree that trust is related to the mental attitude of the trustor, and impacts the trustor's decision-making and selection actions, these concepts should not be conflated.

Many authors (Castelfranchi and Falcone, 2010; Mayer et al., 1995; Delijoo, 2021) emphasize that the evaluation of the trustee's trustworthiness is a mental process internal to the trustor.

Most of the papers devoted to modeling of trust in MAS focus on the problem of trust building: how interaction with other agents influences trust (e.g. (Delijoo, 2021; Sapienza et al., 2022)) without an in-depth discussion of the meaning of particular components of trust and trustworthiness and their relationships.

Following the above, we focus on a foundational conceptual and formal model of the trustworthiness components and show how that can be used for static trust evaluation. We leave the discussion and formal description of the mechanisms for dynamic trustworthiness evaluation for another time.

In the model presented in this paper, like (Mayer et al., 1995; Delijoo, 2021) the concept trustworthiness is build upon three more fundamental concepts:

1. competence
2. benevolence
3. integrity

Below we present definitions of these concepts.

3.1 Competence

Competence, also referred to as ability, is one of the components of trustworthiness. Competence is the potential ability of a trustee to efficiently perform a given task (Delijoo, 2021). Competence is usually domain specific, i.e. a trustee can be competent in one specific domain, but incompetent in another (Poon, 2013). For example, a trustee can be considerably competent in a technical domain but have little social competence. Competence is not a fundamental concept itself as it depends on many other concepts in-

cluding individuals' mental (e.g. personal values and attitudes) and physical characteristics, knowledge and interpersonal skills, amongst others (Delijoo, 2021; Minza, 2019).

For a trustee to be evaluated as competent three main components are needed. The trustee needs a certain set of skills, knowledge and resources within the specific performance domain (Poon, 2013) (Henderson and Cockburn, 1994). Resources are assets which the trustee possesses or has access to. This can be seen as binary, the trustee either has the assets or does not have them. Examples include having access to a working vehicle or being of legal age. Knowledge and skills however, are components of competence which can be acquired by education and experience and can gradually improve. Knowledge does not only entail knowledge within the specific performance domain but also having knowledge of the scope and limits of the task, and being able to come up with an adequate plan to perform the given task (Ruokonen, 2013).

3.2 Benevolence

In (Mayer et al., 1995) benevolence is defined as: *The extent to which a trustee is believed to want to do good to the trustor, aside from an egocentric profit motive.* Benevolence in (Mcknight et al., 2011) is defined as the attitude in which: *One cares about the welfare of the other person and is therefore motivated to act in the other person's interest....does not act opportunistically toward the other...* Both of the above definitions represent similar points of view in which the key is to do good things for the trustor even though it is not necessary perfectly beneficial to the trustee. In terms of trust in machines, (Mcknight et al., 2011) do not use the term benevolence, but its authors substitute it by helpfulness: *The belief that the specific technology provides adequate and responsive help for users.*

3.3 Integrity

The definition of integrity in the context of trustworthiness is relatively consistent across various studies. Following (Barki et al., 2015; Mayer and Davis, 1999; McFall, 1987), our definition of integrity is: *Integrity is the perceived adherence of an individual (trustee) to a consistent set of perceived principles, which are recognized and valued by the trustor.*
Integrity involves maintaining these principles consistently, even in the face of challenges, temptations, or potential personal loss. This adherence must be recognized as being unwavering for reasons deemed right by the trustor, irrespective of whether these principles are universally accepted or approved.

4 A FORMAL CONCEPTUAL MODEL OF TRUSTWORTHINESS

In this section we present our formal conceptual model of trustworthiness that is based on the notions presented in the previous section. First we will formalise competence, then benevolence and last but not least integrity.

4.1 Model of Competence

In pursuit of a widely applicable formalisation, the competence of an agent will be determined by estimating the knowledge, skills and resources of the agent.

It is important to notice that, in order to keep the model at the sufficient level of generality, we do not discuss the representation of concrete skills, knowledge, or resources. Practical implementations would require a concrete, domain dependent, mechanism testing the concrete trustees' skills, knowledge, and resources (e.g. whether the trustee can drive a car).

In some models, competence is assumed to be a binary concept. An agent is deemed competent only if it possesses the requisite knowledge, skills, and resources. Knowledge and skills can gradually improve over time. These two components of competence can therefore be represented as a value on a scale from 0 to 1 and tranformed to binary (if necessary) with the use of thresholds. Resources can be represented as a boolean or a value on a scale from 0 to 1, similar to knowledge and skills.

With respect to representing knowledge, skills and resources, we have to choose between two options. If can either represent all of them as booleans, which would simplify our model and consequently the evaluation function, or we could express them as value (we suggest one between 0 and 1). In that case we could use a threshold function, i.e. a function that maps the values of knowledge, skills and resources to a boolean value representing the trustee to be considered competent or not. The trustor can then choose a certain threshold function for which it deems the knowledge or skills of the trustee sufficient for the task at hand.

If we consider that a task may be build out of smaller sub-tasks, we also have to slice the competence evaluation function in smaller pieces. The overall Competence of a trustee will thus be evaluated as a chain, consisting of the evaluation of the competence for every sub-task. These (sub-)tasks are typically part of a plan of the trustor, that he/she has in mind to be executed by the trustee. Hence in

the formalisation we represent the Goals and their connected sub-plans for that goal, rather than task. Let's assume that an agent, based on its observations and some utility function will select or construct a plan, from a set of plans, each consisting of some sub-plans that may, if completed successfully, results in achieving its goal. In this paper we leave the plan selection for what it is, as we focus on the reliance on other agents to perform certain sub-tasks.

Let:

- $A = \{a_T, a_p, a_q, ...\}$ be a set of agents. Suppose that Agent a_T is a trustor and a_p is a trustee then,

- $P_{a_p} = \{SP_{a_p}^1, SP_{a_p}^2, ..., SP_{a_p}^n\}$ be the plan of agent a_T to be executed by Trustee a_p consisting of a set of subplans that, when completed successfully, achieve the goals that is supposed to be adopted by agent a_p. By P we denote a set of all plans [1]

- Let $g_{a_T}^{SP}$ be a proposition representing an atomic goal of a particular subplan SP of agent a_T in a particular moment of time. By $G_{a_T}^P$ we denote a set of goals in plan P. G denotes a set of all goals[2].

- $C(a_p, SP_{a_p}^i)$ be the competence of agent a_p for subplan $SP_{a_p}^i$

- $\Theta_K(a_p, SP_{a_p}^i)$, $\Theta_S(a_p, SP_{a_p}^i)$ and $\Theta_R(a_p, SP_{a_p}^i)$ be functions returning the knowledge, skills, and resources that agent a_p have for subplan $SP_{a_p}^i$

- $T_K(a_T, SP_{a_p}^i)$, $T_S(a_T, SP_{a_p}^i)$, and $T_R(a_T, SP_{a_p}^i)$ represent the thresholds of knowledge, skills and resources deemed necessary by trustor a_T for subplan $SP_{a_p}^i$

Typically the agent that depends on other agents to execute its plan, will select some sub-plans to be executed to collaborators. We leave out the description for distribution of those sub-tasks over such collaborators here. The plan in the next part of our formalisation is the plan the agent (Trustor) has in mind for the collaborating agent (Trustee).

Definition 1. *Agent a_p (trustee) will be competent to fulfill a plan P_{a_p} which brings about goal $G_{a_T}^{P_{a_p}}$ for agent a_T (trustor) if:*

$$\forall_{SP_{a_p}^i \in P_{a_p}} (C(a_p, SP_{a_p}^i)) \rightarrow C(a_p, G_{a_T}^{P_{a_p}}) \qquad (1)$$

---

[1]Note that we do not present a mechanism of a plan generation (such a mechanism is already implemented in ASC2 (Mohajeri Parizi et al., 2020)), but a mechanism which can control whether a given plan is acceptable for a trustee.

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

when

$$\Theta_K(a_p, SP^i_{a_p}) \geq T_K(a_T, SP^i_{a_p}) \wedge$$

$$\Theta_S(a_p, SP^i_{a_p}) \geq T_S(a_T, SP^i_{a_p}) \wedge$$

$$\Theta_R(a_p, SP^i_{a_p}) \geq T_R(a_T, SP^i_{a_p}) \rightarrow C(a_p, SP^i_{a_p}) \quad (2)$$

Formula 1 can be read as agent a_p is deemed competent by the trustor for the goal $G^P_{a_T}$, if agent a_p is competent for all the subplans of which the goal $G^P_{a_T}$ consists. Formula 2 can be read as agent a_p is deemed competent by the trustor for subplan $SP^i_{a_p}$ if agent a_p has enough knowledge, skills and resources for the subplan, i.e. that all values for Knowledge, Skills and Resources are beyond the given thresholds.

4.2 The Model of Benevolence

We are going to use *values* as the central concept allowing for representation of the agents' goals. The concept of value used in our model was as introduced in (Zurek, 2017) and later developed in (Wyner and Zurek, 2024), where value is defined as an abstract (trans-situational) concept which allows for the estimation of a particular state of affairs and influences one's behavior. The key assumption is that every goal, understood as 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 model of benevolence was introduced in (Zurek et al., 2025)). Moreover, since our model is constructed to be implemented with an AS2/AgentSpeak framework which allows for building complex plans including a number of recursively invoked goals, we assume that there is also the possibility to evaluate a whole plan including a number of goals. In order to make such a comparison, we introduce basic concepts:

- Suppose a set of values: $V = \{v_a, v_b, v_c, ...\}$.

- Let $\Phi : A \times V \times 2^G \rightarrow \langle 0; 1 \rangle$ be a function returning the level of satisfaction of value from set V by a subset of G in the eyes of a given agent. For example, by $\Phi_{v_a}(a_p, \{g^{SP_i}_{a_p}, g^{SP_2}_{a_p}\}) = 0.5$ we denote that the joint level of satisfaction of value v_a by goals $g^{SP_1}_{a_p}, g^{SP_2}_{a_p}$ in the eyes of agent a_p is 0.5. If a particular plan P_{a_p} results in achieving two goals $g^{SP_1}_{a_p}, g^{SP_2}_{a_p}$, then $\Phi_{v_a}(a_p, G^{P_{a_p}}_{a_p}) = \Phi_{v_a}(a_p, \{g^S_{a_p}, g^T_{a_p}\})$.

- We say that a new plan P'_{a_p} of agent a_p demotes value v_a w.r.t. initial plan P_{a_p} if $\Phi_{v_a}(a_p, G^{P'_{a_p}}_{a_p}) < \Phi_{v_a}(a_p, G^{P_{a_p}}_{a_q})$, is neutral w.r.t. v_a

if $\Phi_{v_a}(a_p, G^{P'_{a_p}}_{a_q}) = \Phi_{v_a}(a_p, G^{P_{a_p}}_{a_q})$, and promotes v_a otherwise.

What is important here is that the agent (trustee) may have a different attitude towards values; some of them are not of great importance and they can be easy sacrificed, while some of them are crucial for the agent and he is not going to demote them. This leads to the conclusion that a trustee's willingness to demote his goals, in order to support the trustor's ones, is the matter of the possibility to demote a set of values, each of which can have different threshold. On the basis of the above:

Definition 2. *Let $\Gamma : A \times V \rightarrow \langle 0; 1 \rangle$ 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 $a \in A$*

Since our work aims at evaluating trustworthiness, we should not evaluate the actual benevolence of a trustee if possible at all, but rather the perceived benevolence in a trustor's mind. Therefore, we will index the Γ function w.r.t. the agent (trustor) who make the benevolence evaluation. 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 3. *Let $G^{P_{a_p}}_{a_p}$ be a set of goals of a initial plan of agent a_p. A new plan $P^N_{a_p}$ which fulfills goal $G^{P^N_{a_p}}_{a_p}$ will be acceptable for agent a_p and agent a_p will be sufficiently benevolent for adopting this plan in a view of trustor a_T, which we denote by $BEN(a_p, P^N_{a_p})$ if:*

$$\forall_{v_x \in V}(\Phi_{v_x}(a_p, G^{P_{a_p}}_{a_q}) < (\Gamma_{a_T}(a_p, v_x) +$$

$$\Phi_{v_x}(a_p, G^{P^N_{a_p}}_{a_p}))) \rightarrow BEN(a_p, P^N_{a_p}) \quad (3)$$

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

4.3 Model of Integrity

The concept integrity as defined before is build upon two other concepts, principles and intentions. Therefore we formalise Integrity as follows:
Let *PR* represent the set of these principles. Although it lacks a common understanding what principles are in literature (Dworkin, 1978; Alexy, 2003), undoubtedly, they are something more generic than usually

more specific rules. In this paper we adhere to the concept principle as presented in (Zurek et al., 2022), where a principle is represented as a minimal acceptable level of satisfaction of a particular (societal) value.

The trustor may assume adherence to a number of principles and the strength of that adherence can be expressed as a value between 0 and 1, where 0 means no adherence and 1 fully adhering to the principle.

The distance between the perceived principles and the perceived intentions, D, is calculated using a Euclidean distance formula, $D(P)$.

To ensure that the distance is always normalized between 0 and 1, it is divided by the maximum possible distance, D_{max}, which represents the maximum Euclidean distance between any two points. This normalization produces $D_{normalized}$.

Finally, the adherence to the trustor's principles or integrity, α, is defined as 1 minus this normalized distance. Therefore, an integrity score close to 1 indicates high adherence to the Trustor's principles, while a score close to 0 indicates low adherence. This way, integrity is mathematically represented as the inverse of the normalized distance between the set of principles that are not in conflict with their set of intentions. Let:

- D be a variable representing distance;
- $PR_{a_T} = \{pr_{a_{a_T}}, pr_{b_{a_T}}, \ldots, pr_{n_{a_T}}\}$ be a set of principles that are deemed acceptable by the trustor;
- Following (Zurek et al., 2022), we assume that principle can be expressed by a desired level of satisfaction of a values. By $V(PR_{a_T}) = \{v_a(pr_{a_{a_T}}), v_b(pr_{b_{a_T}}), \ldots, v_n(pr_{n_{a_T}})\}$ we denote a set of desirable, for a trustor, levels of satisfaction of values. There may be some tension for an agent adhering to some principles that result in promoting some (societal) values and the extend to which accomplishing some (sub) goals promote them. We express this by a Tension function, that can be interpreted as the inverse of integrity:
  $T_a^{a_p} = v_a(pr_{a_{a_T}}) - \Phi_{v_a}(a_p, G_{a_p}^{P_{a_p}})$;
- Let T^{a_p} be a set of all $T_a^{a_p}$ of agent a_p;

D_{max} is maximum possible Euclidean distance between any two points in an n-dimensional space (where n is a number of values), where each dimension's range is from 0 to 1:

$$D_{max} = \sum_{k=1}^{n} 1 \qquad (4)$$

$D(T^{a_p})$ is the Euclidean distance between two points indicating the relevance of the concordant principle:

$$D(T^{a_p}) = \sqrt{\sum_{k=1}^{n} (T_k^{a_p})^2} \qquad (5)$$

This normalization ensures that $D_{normalized}$ is always between 0 and 1.

$$D_{normalized} = \frac{D(V(PR))}{D_{max}} \qquad (6)$$

This inversion converts the distance to a similarity score: a value close to 1 indicates low distance (high similarity or adherence), and a value close to 0 indicates high distance (low similarity or adherence). On the basis of the above we may define the formal representation of integrity:

Definition 4. *The assumed Integrity of agent a_p (the trustee) by agent a_T is expressed as:*

$$I_{a_p} = 1 - D(T^{a_p}) \qquad (7)$$

If by $T_I(a_T, P_{a_p})$ we assume the trustor's threshold for acceptable integrity of potential trustee, then:

Definition 5. *Trustor a_T will perceive potential trustee a_p as satisfactory integer for fulfilling plan P_{a_p}, which we denote by $I(a_p, P_{a_p})$ if:*

$$I_{a_p} \geq T_I(a_T, P_{a_p}) \rightarrow I(a_p, P_{a_p}) \qquad (8)$$

In other words, trustee will be perceived as integer if the distance of values of trustor and trustee will be less than certain threshold.

4.4 Evaluation of Trustworthiness

In this section we integrate the above models to introduce the joint evaluation of trustworthiness. Although we are going to join the above models together, we are not going to boil everything down to one number. In our opinion, decision whom to trust, takes into consideration the evaluation of all the components of trustworthiness independently.

On the basis of such an intuition we assume that the evaluation of trustworthiness should be represented by all the components of trustworthiness:

Definition 6. *The trustworthiness of agent a_p in the eyes of trustor a_T can be represented as triple:*

$$TR_{a_p}^{a_T} = \langle C(a_p, G_{a_T}^{P_{a_p}}), BEN(a_p, P_{a_p}), I(a_p, P_{a_p}) \rangle \quad (9)$$

4.5 Experimental Implementation

We model intentional agents via the *belief-desire-intention* (BDI) model (Rao and Georgeff, 1995). In practice, BDI agents also include concepts of *goals* and *plans*. Goals are concrete desires, plans are abstract specifications for achieving a goal,

and intentions then become commitments towards plans. Our implementation was made with the use of AgentScript/ASC2 (Mohajeri Parizi et al., 2020) language. The implementation with the description of the example scenario can be found on the project's github: https://github.com/basten-leeftink/PaperTrustworthiness/.

5 DISCUSSION AND CONCLUSIONS

Trust is considered as one of the most important elements shaping the modern society. Our project aims at exploring the nature of trust by reproducing the the real social relations between agents within Multi-Agent system. In order to fulfill the aim, we have to model and implement all the necessary components of trust into MAS agent both on the micro level of single agent and macro level of the society of agents. In this paper we focused on the representation of the trustworthiness of a potential trustee. Our main contribution is in extending, clarifying, and modeling the concept of trustworthiness for MAS. In most of the existing models (e.g (Delijoo, 2021; Sapienza et al., 2022)) trustworthiness and trust is represented by a number. In our opinion, these terms has much more complex character, and to represent trustworthiness in a more accurate way, we decomposed this into its three main dimensions: competence, benevolence, and Integrity.

Following BDI architecture, plans of agents are recursively divided into subplans. Every agent to be satisfactory competent to perform a delegated plan must be satisfactory competent for every subplan. In order to represent this concept we divided the concept of competence into its sub-components: knowledge, skills, and resources and introduced the trustor's thresholds for each of these sub-concepts. These thresholds represent the least acceptable level of each of these concepts.

Our key observation concerning benevolence is that it should be presented in the light of the agent's (trustee's) goal: how much he/she can sacrifice with respect to his/her initial plans. In order to do that, we introduce the notion of goal and values as a background of the agent's goal. In our work we adopt the concept of goal from (Zurek, 2017; Wyner and Zurek, 2023) in which goal is represented by a set of the levels of satisfaction of different values. On the basis of that, we can also observe that the agent may have different willingness to sacrifice different values. Following that, our model assumes that benevolence of an agent can be understood as a set of levels of the acceptable (for a given agent) demotion of different values.

The concept of integrity is particularly interesting. Similar to benevolence, we use values to represent agent's integrity. We are using a weighted inverted Euclidean distance formula. The intuition is that understood as a distance between values of trustor and trustee. This paper has shown how the values of trustee can be quantified and compared against a set of predefined principles of trustor.

One of the most important element of our trustworthiness model is that we do not reduce everything to a single value, but since trustworthiness is a multidimensional concept, we assume that it is expressed by a tuple of its three main components. For the sake of this paper we assumed that each component has a binary character (trustee is either competent or not, benevolent or not, etc.) but for more complex decision making processes, competence and integrity can be represented without thresholds just as the levels satisfaction of these parameters. Such an assumption of the multidimensional character of trustworthiness allows for much more informed decision of the agent on one hand (different types of delegations and actions put a different role importance on every component), and allows for placing trust on the agents which are not fully trustworthy (what often happens in a real life situations) on the other hand. Note that the separation of the trustworthiness model from the decision making process, allows for using our model in various types of decision making mechanism (including knowledge-base as well as machine-learning paradigm)

One of the aims of our project, was to keep the model flexible enough to be implemented in various systems with different purposes, including research on trust or the research on socio-technical normative systems. In order to fulfil this requirement, the model has been created in a modular way, by which we mean that each part can be implemented separately (for example, only benevolence without competence and integrity evaluation). Moreover, it can be used for a binary evaluation (someone is either trustworthy or not) or for a gradual one (competence or integrity can be represented as a number). Such a construction of the model, gives flexibility necessary for different tasks.

Although our model is, to our knowledge, the most comprehensive representation of trustworthiness evaluation (see section 3 for deeper discussion), it still has some limitations. For example, it does not take into consideration personal sympathies of a trustor. This is, however, a matter of balance between complexity and efficiency. More complex models require much more background knowledge and data, and are significantly more computationally demanding.

Future research will include modelling the dynamics of the trustworthiness evaluation, in particular the mechanisms of the influence of external signals, like experience or communication with other agents, on the changes of the evaluation of potential trustees trustworthiness. This will be a basis of the design of the decision making mechanism (who to trust?) and the experimental verification of our model.

REFERENCES

Alexy, R. (2003). On balancing and subsumption. a structural comparison. *Ratio Juris*, 16(4):433–449.

Barki, H., Robert, J., and Dulipovici, A. (2015). Reconceptualizing trust: A non-linear boolean model. *Information & Management*, 52(4):483–495.

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

Chen, S.-H., Chie, B.-T., and Zhang, T. (2015). Network-based trust games: An agent-based model. *Journal of Artificial Societies and Social Simulation*, 18(3):5.

Delijoo, A. (2021). *Computational trust models for collaborative network orchestration*. PhD thesis, University of AMsterdam.

Dworkin, R. (1978). *Taking Rights Seriously. New Impression with a Reply to Critics*. Duckworth.

Frey, V. and Martinez, J. (2024). Interpersonal trust modelling through multi-agent reinforcement learning. *Cognitive Systems Research*, 83:101157.

Fung, H. L., Darvariu, V.-A., Hailes, S., and Musolesi, M. (2022). Trust-based consensus in multi-agent reinforcement learning systems. *ArXiv*, abs/2205.12880.

Henderson, R. and Cockburn, I. (1994). Measuring competence? exploring firm effects in pharmaceutical research. *Strategic management journal*, 15(S1):63–84.

Jaffry, S. W. and Treur, J. (2013). *Agent-Based and Population-Based Modeling of Trust Dynamics*, pages 124–151. Springer Berlin Heidelberg, Berlin, Heidelberg.

Mayer, R. and Davis, J. H. (1999). The effect of the performance appraisal system on trust for management: A field quasi-experiment. *Journal of Applied Psychology*, 84:123–136.

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

McFall, L. (1987). Integrity. *Ethics*, 98(1):5–20.

Mcknight, D. H., Carter, M., Thatcher, J. B., and Clay, P. F. (2011). Trust in a specific technology: An investigation of its components and measures. *ACM Trans. Manage. Inf. Syst.*, 2(2).

Minza, M. (2019). Benevolence, competency, and integrity: Which one is more influential on trust in friendships? *Jurnal Psikologi Vol*, 18(1):91–105.

Mohajeri Parizi, M., Sileno, G., van Engers, T., and Klous, S. (2020). Run, agent, run! architecture and benchmarking of actor-based agents. In *Proceedings of the 10th ACM SIGPLAN International Workshop on Programming Based on Actors, Agents, and Decentralized Control*, AGERE 2020, page 11–20, New York, NY, USA. Association for Computing Machinery.

Nobandegani, A. S., Rish, I., and Shultz, T. R. (2023). Towards machines that trust: Ai agents learn to trust in the trust game. *ArXiv*, abs/2312.12868.

Parsons, S., Sklar, E., and McBurney, P. (2012). Using argumentation to reason with and about trust. In McBurney, P., Parsons, S., and Rahwan, I., editors, *Argumentation in Multi-Agent Systems*, pages 194–212, Berlin, Heidelberg. Springer Berlin Heidelberg.

Poon, J. M. (2013). Effects of benevolence, integrity, and ability on trust-in-supervisor. *Employee Relations*, 35(4):396–407.

Rao, A. S. and Georgeff, M. P. (1995). Bdi agents: From theory to practice. In *Proceedings of the First International Conference On Multi-Agent Systems (ICMAS-95)*, pages 312–319.

Ruokonen, F. (2013). Trust, trustworthiness, and responsibility. In *Trust*, pages 1–14. Brill.

Sapienza, A., Cantucci, F., and Falcone, R. (2022). Modeling interaction in human-machine systems: A trust and trustworthiness approach. *Automation*, 3(2):242–257.

Tykhonov, D., Jonker, C., Meijer, S., and Verwaart, D. (2008). Agent-based simulation of the trust and tracing game for supply chains and networks. *Journal of Artificial Societies and Social Simulation*, 11(3):1–32.

Wyner, A. and Zurek, T. (2024). Towards a formalisation of motivated reasoning and the roots of conflict. In Osman, N. and Steels, L., editors, *Value Engineering in Artificial Intelligence*, pages 28–45, Cham. Springer Nature Switzerland.

Wyner, A. Z. and Zurek, T. (2023). On legal teleological reasoning. In Sileno, G., Spanakis, J., and van Dijck, G., editors, *Legal Knowledge and Information Systems - JURIX 2023: The Thirty-sixth Annual Conference, Maastricht, The Netherlands, 18-20 December 2023*, volume 379 of *Frontiers in Artificial Intelligence and Applications*, pages 83–88. IOS Press.

Zurek, T. (2017). Goals, values, and reasoning. *Expert Systems with Applications*, 71:442 – 456.

Zurek, T., Araszkiewicz, M., and Stachura-Zurek, D. (2022). Reasoning with principles. *Expert Systems with Applications*, 210:118496.

Zurek, T., Wyner, A., and van Engers, T. (2025). The model of benevolence for trust in multi-agent system. To appear in: Proceedings of 18th KES International Conference, KES-AMSTA 2024, June 2024.