Trust should correspond to Trustworthiness: a Formalization of Appropriate Mutual Trust in Human-Agent Teams

Carolina Centeio Jorge[1]
C.Jorge@tudelft.nl

Siddharth Mehrotra[1]
S.Mehrotra@tudelft.nl

Catholijn M. Jonker[1,2]
C.M.Jonker@tudelft.nl

Myrthe L. Tielman[1]
M.L.Tielman@tudelft.nl

[1] Department of Intelligent Systems, TU Delft
[2] Leiden Institute of Advanced Computer Science, Leiden University

Abstract

In human-agent teams, how one teammate trusts another teammate should correspond to the latter's actual trustworthiness, creating what we would call appropriate mutual trust. Although this sounds obvious, the notion of appropriate mutual trust for human-agent teamwork lacks a formal definition. In this article, we propose a formalization which represents trust as a belief about trustworthiness. Then, we address mutual trust, and pose that agents can use beliefs about trustworthiness to represent how they trust their human teammates, as well as to reason about how their human teammates trust them. This gives us a formalization with nested beliefs about beliefs of trustworthiness. Next, we highlight that mutual trust should also be appropriate, where we define appropriate trust in an agent as the trust which corresponds directly to that agent's trustworthiness. Finally, we explore how agents can define their own trustworthiness, using the concepts of ability, benevolence and integrity. This formalization of appropriate mutual trust can form the base for developing agents which can promote such trust.

1 Introduction

Artificial agents are becoming more intelligent and able to execute relevant tasks for our daily lives, including work environments, home assistance, battlefield and crisis response [LSW18]. For some of these tasks, humans and artificial agents should learn to cooperate, coordinate and collaborate, forming *human-agent teams*. A key driver for achieving effective teamwork is *mutual trust* [SSB05], i.e., teammates should trust each other. In particular, we consider that *appropriate* mutual trust is a fundamental property in effective human-agent teamwork. We take appropriate to mean that a human's trust in an agent should correspond to that agent's trustworthiness, and an agent's trust in a human should correspond to the human's trustworthiness. To achieve

In: R. Falcone, J. Zhang, and D. Wang (eds.): Proceedings of the 22nd International Workshop on Trust in Agent Societies, London, UK on May 3-7, 2021, published at http://ceur-ws.org

this form of mutual trust, we should develop artificial agents that are able reason about and promote appropriate mutual trust. To this end, we propose a formalization of the beliefs and concepts involved in appropriate mutual trust, in the context of human-agent teamwork.

Appropriate trust in teams happens when one teammate's trust towards another teammate corresponds to the latter's actual trustworthiness. When there is appropriate trust, there is no under-trust (leading to under-reliance) or over-trust (leading to over-compliance), in human-agent teams [LLS20], which minimizes negative performance outcomes [OSPJ13]. Trust corresponds to how we expect the other to perform a particular action, whereas trustworthiness is how others will actually perform that action. For example, if an agent x trusts another agent y to execute a task (e.g., driving a car) which requires skills that y does not have, agent x over-trusted agent y and the consequences can be negative and even disastrous (e.g., car accident). On the other hand, if agent x does not trust agent y to execute a task (e.g., driving a car) and agent y is perfectly capable of successfully executing the task, agent x is under-trusting agent y which can also negatively affect team effectiveness (e.g., walking instead). In particular, when x is a human and y is an artificial agent, and trust is not appropriate, this will lead to disuse or misuse of technology [LSW18]. Thus, a dyadic relationship between a human and an artificial agent in a human-agent team should be designed in such a way that it supports 1) appropriate trust from the human towards the agent and 2) appropriate trust of the agent towards the human.

We approach trust from a functional perspective, in which trust is a relational construct between the trustor x, the trustee y, about a defined (more or less specialized) task (τ), as in [FPVC13]. Particularly, we propose that trust is one agent's *perception of the trustworthiness* of another, meaning that how much x trusts y depends on how trustworthy (e.g., y's ability and willingness to drive) x believes y is. This means, x appropriately trusts y when x's belief in y's trustworthiness actually corresponds to y's trustworthiness. As such, for an artificial agent to appropriately trust a human, we first need to understand a human's trustworthiness, and secondly how the agent's beliefs about human's trustworthiness could be formed. Moreover, artificial agents should promote and elicit appropriate trust in themselves. So if y wants x to trust y more (or less), y has to make x believe y is more (or less) trustworthy [FDA14] (e.g. by showing off y's driving skills, or showing third party recommendations, etc). To do this, agents need to know whether their human teammate's trust in them corresponds to their actual (own) trustworthiness. Moreover, this means that agents need to understand how humans form beliefs of trust, and how the agents assess their own trustworthiness (which is also called a belief in the well-known BDI architectures [RG95]). The detailed dynamics and formation of these beliefs, see e.g. [BJTT07, HLHV09], is out of the scope of this paper.

Trust has been vastly explored in the context of human-human interaction, with well-known contributions such as *ABI* [MDS95] model, for organizational behaviour. In particular, trust in human teams has been recently explored in contexts such as virtual teams [BHHH20], sports [HJW19], and university group projects [NPW18]. Furthermore, in Multi-agent Systems (MAS), trust has been used as a security and control mechanism, to protect agents from not knowing other agents' code of conduct [SMV13]. Among others, we can find a formalization for trust and reputation, see e.g. [HLHV09], ways of categorizing agents to explain internal qualities (krypta) with their observable signs (manifesta) in order to promote trust, see e.g. [FPVC13, BNS13], and, more recently, models for assessing agent's trust based on human values, see e.g. [CNGD19, CMP17]. Similarly, trust in human-agent interaction has been gaining increasing attention, including the dynamics of human's trust towards technology, see e.g. [Win18, NWL$^+$20], how agents can assess and promote appropriate trust in humans, see e.g. [FDA16, ASKL19, GY20, NGP$^+$20], and the role of (appropriate) trust in human-agent teams, see e.g. [UG20, LLS20, SPG$^+$21, VMP$^+$20].

Essentially, we can find in the literature 1) how humans trust humans, 2) how agents can trust other agents, 3) how humans trust artificial agents, and 4) how artificial agents can calibrate this trust with certain actions. However, we found literature on trust from the perspective of an agent towards a human to be scarce (we extend this discussion in Section 3). In particular, there is a research gap in studying 1) how agents can trust humans, 2) how humans' and agents' trustworthiness and trust in each other are related and influence each other, including, 3) how the manipulation of an agent's own trustworthiness can manipulate a human's trust in that agent. We propose that to support studies into these questions, we need to to formalize beliefs about trustworthiness (i.e. trust), and beliefs about trust (i.e., beliefs of beliefs of trustworthiness) and their dependencies. In this paper, we discuss why we need nested definition of beliefs to understand trustworthiness and trust of humans and artificial agents, in the context of human-agent teamwork. We relate these definitions to prior work and make inferences on how we can actually utilize this. For that, we propose a formalization for:

1. Trust as a belief of trustworthiness;

2. Trustworthiness of the human towards an agent;

3. Agent's belief in human's trustworthiness towards an agent;

4. Agent's belief in human's trust towards an agent;

5. Agent's trustworthiness calibration (through the agent's belief in its own trustworthiness) for appropriate human's trust.

The main contribution of this paper is a formalization of appropriate mutual trust in human-agent dyadic relationships for supporting the design of human-agent teams. The remainder of this paper is organized as follows: Section 2 explores the concept of trust as a belief of (directed) trustworthiness, Section 3 formalizes the agent's trust in the human in the context of human-agent teams and, Section 4 explores the belief formation on human's trust with the goal of appropriate trust. Finally, we discuss an example in Section 5, limitations and future work in Section 6, and conclude in Section 7.

2 Trust as a belief of Trustworthiness

Humans rely on trust on a daily basis, every time we need to interact with, delegate to or rely on the intention of another individual, group or thing [URO13b]. In a dyadic relation between two *cognitive agents* [CF00] (artificial or human), trust involves two parties, the *trustor* and the *trustee*, and an action (trusted by the trustor to the trustee) that affects a goal (of the trustor) [CF10]. *Trust* and *trustworthiness* are two similar concepts, which are related, but distinct from each other. While trustworthiness, the characteristic that someone is to be trusted, is an objective property of the trustee, trust is a subjective attitude of the trustor, which involves the *perceived* trustworthiness of the trustee. This implies that the trustor must have a "theory of the mind" of the trustee, which may include personality, shared values, morality, goodwill, etc [CF00]. Trust is an aspect of relationships and, as such, can only be viewed in the context of individuals and their relationships [SMD07]. As an example, let's imagine that a cognitive agent y (artificial or human) drives well and is trustworthy regarding driving tasks. For another cognitive agent x to trust agent y for a driving task, agent x has to *believe* that agent y is trustworthy for this task. This corresponds to the concept that any changeable notion that an agent has about the world is a *belief* that agent has. In this, we follow the Belief-Desire-Intention (BDI) architecture for agent [RG95]. This being said, we propose that trust of agent x in agent y, T, is a *belief* of x (trustor), $\mathcal{B}_x$, about y's (trustee's) trustworthiness, $\mathcal{TW}_y$, meaning that:

$$T(x, y) = \mathcal{B}_x(\mathcal{TW}_y) \tag{1}$$

Accordingly, in order to understand trust, we first need to understand trustworthiness, and secondly how beliefs about trustworthiness are formed. Trustworthiness is a complex concept, and following the literature it can consist of a set of dimensions that range from the trustee's competence to its intentions [Gri05]. How an entity can be considered trustworthy is not a trivial question, and is context-dependent, as well as dependent on the nature of the trustee [SPG+21, HLHV09]. When considering human trustworthiness in organizational behaviour, *Ability, Benevolence and Integrity* (*ABI*) model [MDS95] is often employed. Similarly, we can consider the most consensual dimensions of trust (perceived trustworthiness) in technology as being *Performance, Process and Purpose* [LS04], but understanding how these can model trustworthiness itself would require further study. When talking of artificial agents and societies, for example, we can consider beliefs such as *Willingness*, *Competence* and *Dependence* to estimate the trustworthiness of another cognitive agent [CF10]. Finally, trustworthiness or its dimensions can be affected by external factors, which are contextual conditions determining the situation in which the task is executed [FPVC13], such as environmental configuration, emotional state, workload, etc.

As mentioned previously in this section, trust varies with the person and across relationships. We will illustrate this using the *ABI* model of trust [MDS95], which has been widely used to study trustworthiness. The authors define trust as *"the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party"*. In this model of trust, trustworthiness is defined as "the extent to which an actor has the ability to execute relevant tasks, demonstrates integrity, and is benevolent towards fellow team members" [VMP+20]. Furthermore, in the paper the authors define *Ability*, *Benevolence* and *Integrity* as follows:

1. **Ability**: Ability is that group of skills, competencies, and characteristics that enable a party to have influence within some specific domain.

2. **Benevolence**: Benevolence is the extent to which a trustee is believed want to do good to the trustor, aside from an egocentric profit motive. Benevolence suggests that the trustee has some attachment to the trustor.

3. **Integrity**: The relationship between integrity and trust involves the trustor's perception that the trustee adheres to a set of principles that the trustor finds acceptable.

We can see that although *Ability* depends only on the trustee, both *Benevolence* and *Integrity* depend on both the trustor and the trustee. Even though trustworthiness is a characteristic of a trustee, this characteristic will differ per trustor. Following, both trust and trustworthiness depend on the two cognitive agents (artificial or human), trustor and trustee (x and y), that compose the dyadic relationship. Thus, we stipulate that we need to define the trust of an agent x in agent y as the belief $\mathcal{B}$ of agent x regarding the trustworthiness of y with respect to x, adapting Expression 1 to:

$$T(x, y) = \mathcal{B}_x(\mathcal{TW}_y(x)) \tag{2}$$

3 Artificial agents trusting humans

We view trust as a directional and dyadic relationship, which, within human-agent teams, is composed of either humans or artificial agents. This adds complexity since artificial agent and human's beliefs are often constructed in very different ways, which means beliefs about trustworthiness are so as well. Moreover, estimating the trustworthiness properties of someone who also thinks and operates differently from you will always be more challenging. There are contributions on how an artificial agent can detect that a situation requires trust [WA11, WRH18] and also how an artificial agent can detect whether a human is being trustworthy, based on episodic memory [VPCC19] and social cues [SW19]. Although we consider this research highly contributory to the field of artificial agent's trust towards humans, at this stage it is still preliminary. Making artificial agents able to detect under which situations they could use trust and when they can trust a human, based on social cues and memory, is important. However, enabling them to understand human trustworthiness and its dimensions can lead to another level of human-agent understanding and team effectiveness. For this reason, in this section we propose a formalization that will enable the agent to reason about the human's trustworthiness.

As we have mentioned before, in human-agent teaming, it is important that 1) the agent appropriately trusts the human and 2) the human appropriately trusts the agent, but it is also important that 3) the agent has a *belief* about whether the human appropriately trusts the agent, and that 4) the human *believes* the agent appropriately trusts the human, and why. In this paper we will focus on the cases 1 and 3, since we are addressing trust from the agent's perspective as we can only directly manipulate the artificial agent's beliefs. Fig. 1 schematizes a dyadic human-agent relationship, including these four concepts.

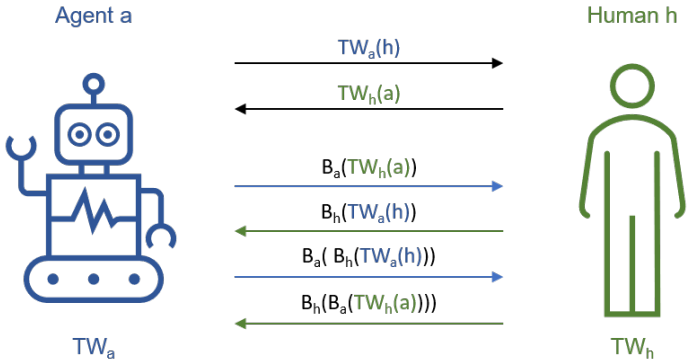

Figure 1: Human-Agent dyadic trust.

As an example, let us again consider the task of driving a car. Inspired by [MdS20], let's imagine a dual-mode vehicle, which can be driven both by an artificial agent or by a human. The default setting is the human driving according to the agent's instructions, but the agent takes over when it recognizes dangerous situations. Although it may be counter-intuitive, we need the agent to *trust* the human to drive safely (their joint goal), while complying to societal ethics, so that it knows when to take over. Both human and artificial agent have their

property of trustworthiness (e.g., regarding driving), $\mathcal{TW}_h$ and $\mathcal{TW}_a$, respectively. However, when applying this trustworthiness in a dyadic relationship, this trustee's property becomes dependent on the trustor, as formalized in section 2. In this example, we will have the trustworthiness of the agent a, given a human h, $\mathcal{TW}_a(h)$, and the trustworthiness of the human h given an artificial agent a, $\mathcal{TW}_h(a)$. In practical terms, this means that the way the human is going to follow the agent's instructions, may vary according to the agent that is helping (e.g., depending on whether the human relies on this particular agent's knowledge/intelligence). Moreover, we have the trust of the artificial agent in the human, meaning the agent's belief on human's trustworthiness, $T(a,h) = \mathcal{B}_a(\mathcal{TW}_h(a))$ (from Expression 2), and the trust of the human in the agent, which is the human's belief on agent's trustworthiness $T(h,a) = \mathcal{B}_h(\mathcal{TW}_a(h))$. The trust of the artificial agent in the human ($T(a,h)$) is what the agent believes that the human will do if the agent gives the human a certain instruction.

An agent, to be able to promote and elicit appropriate trust (from the human towards the agent), does not only need to reason with beliefs about human's trustworthiness, but also with beliefs about human's trust (estimating whether the human trusts the agent). What's more, we identify that we may also need beliefs about trust when appropriately estimating human's trustworthiness. This being said, in the dual-mode vehicle example, can the agent trust the human to follow an instruction, if the human does not trust that agent? Considering again the *ABI* trustworthiness model mentioned in the previous section, we believe that if a trustee trusts a trustor, this is a sign of their benevolence towards the trustor, which in turn would increase the trustee's trustworthiness to that trustor. Thus, in order to trust the human teammate, the agent should estimate the human's trust in the agent. In order to estimate $\mathcal{B}_a(\mathcal{TW}_h(a))$, we may also need the agent's belief in human's trust in the agent, i.e., $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h)))$. Following the example, for the agent to trust the human to follow an instruction, the agent needs to believe that the human trusts the agent (e.g., the human relies on this particular agent's knowledge/intelligence).

The ultimate goal of the agent is to appropriately trust the human. So, when estimating whether it can trust its human teammate to follow an instruction, the *agent*'s trust in the human should correspond to the actual human's trustworthiness (e.g., to what actually the human can and/or wants to do), i.e.,

$$T(a,h) \equiv \mathcal{B}_a(\mathcal{TW}_h(a)) \equiv \mathcal{TW}_h(a) \tag{3}$$

which requires that the agent also accurately estimates the human's trust in the agent, $T(h,a) \equiv \mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h)))$. The *human*'s trust in the agent, on the other hand, is the belief of the human in the agent's trustworthiness, $\mathcal{B}_h(\mathcal{TW}_a(h))$, and should correspond to the agent's actual trustworthiness ($\mathcal{TW}_a(h)$), i.e.,

$$T(h,a) \equiv \mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h))) \equiv \mathcal{B}_h(\mathcal{TW}_a(h)) \equiv \mathcal{TW}_a(h) \tag{4}$$

Lastly, since the nested concepts presented on Expression 4 are based on $\mathcal{TW}_a(h)$, this means that we may be able to calibrate human's trust in the agent ($T(h,a)$), by manipulating $\mathcal{TW}_a(h)$ through the accurate belief of the agent's own trustworthiness. This means that if the agent is aware of its own trustworthiness, meaning that if the agent's belief in agent's trustworthiness corresponds to actual agent's trustworthiness, i.e.,

$$B_a(\mathcal{TW}_a(h)) \equiv \mathcal{TW}_a(h) \tag{5}$$

the agent may be able to alter its own trustworthiness (or simply how it lets the human perceive it) and, consequently, calibrate human's trust. In our example, the agent might understand that it is not being perceived as intelligent, and start justify its instructions, possibly leading the human to trust it more. These expressions are not trivial to achieve since they rely on having a criteria for appropriate level of belief, which is something difficult both to formalize and measure. However, we believe they show the direction in which these beliefs should be formed, and what should be kept in mind while designing these systems.

To showcase how we can use this formalization, we can consider *ABI* trustworthiness model once more. We consider the trustworthiness of the human to be the weighted sum of their ability, integrity and benevolence towards a specified task, in a certain environment. Similarly, we consider agent's trust in the human to be the weighted sum of the agent's beliefs in the ability, benevolence and integrity of the human, towards a specified task (τ), in a certain environment (ϵ) (both task and environment as in [FPVC13]). The belief in *ability* (*Ab*) of the human takes into account the task τ and environment ϵ. The belief in human's *benevolence* (*Ben*) however, besides the task and environment, also takes into account the agent (since it is directed, as seen in Section 2). Benevolence may also, among other things, implicitly use the belief of the human's trust in the agent, $\mathcal{B}_a(T(h,a))$ (which, as previously discussed, can be expressed as $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h)))$. Finally, the belief in human's *integrity* (I) depends on the agent, task and environment. Thus,

$$\mathcal{B}_a(\mathcal{TW}_h(a,\tau,\epsilon)) = W \cdot [\mathcal{B}_a(Ab_h(\tau,\epsilon))\ ,\ \mathcal{B}_a(Ben_h(a,\tau,\epsilon)\ ,\ \mathcal{B}_a(I_h(a,\tau,\epsilon))] \tag{6}$$

where W is a weight vector.

4 Humans trusting artificial agents

We now shift our focus to understanding how humans trust agents. As mentioned before, it is important that human's trust in the artificial agent is appropriate. The challenge is to ensure that humans tune their trust towards the agent, since we do not have control over the human. However, leveraging on the idea that agents reflect about their own trustworthiness, we may be able to influence humans to appropriately fine-tune their trust in them. With the information regarding the agent's trustworthiness, human teammates can adapt to the qualities and limitations of the agent and, consequently, adjust the utilization of the agent accordingly. Without this knowledge, however, it would prove difficult to coordinate within a task-environment given an unpredictable agent teammate. For example, let us again consider the task of driving a car. Considering that the agent reflects about its own trustworthiness regarding its ability and willingness to drive the vehicle, the agent may then influence the human teammate to adapt to the agent's strengths and weaknesses (fine-tuning human's trust in the agent).

4.1 Understanding appropriate trust

We posit that how trustworthy an agent is for a human and how a human trusts the agent (human's belief in agent's trustworthiness) should be similar to get appropriate trust. If the belief of an agent in their own trustworthiness towards human is different from their belief of human's trustworthiness towards them then we come closer to under-trust $T(a,h)\downarrow$ or over-trust $T(a,h)\uparrow$ *i.e.*

$$\mathcal{B}_a(\mathcal{TW}_a(h)) \ > \ \mathcal{B}_a(\mathcal{TW}_h(a)) \rightarrow T(a,h)\uparrow \tag{7}$$

$$\mathcal{B}_a(\mathcal{TW}_a(h)) \ < \ \mathcal{B}_a(\mathcal{TW}_h(a)) \rightarrow T(a,h)\downarrow \tag{8}$$

Therefore, to avoid such situations, the agent's belief in their own trustworthiness should match with their belief about the belief of human's trustworthiness in them. This will result in eliciting appropriate trust in a human from an agent perspective. Next, we turn to understanding how to develop appropriate trust. Researchers have defined appropriate trust in an artificial agent in different forms. A common outline of many definitions is related to the capability or ability of an agent. Here, appropriate trust is the alignment between the perceived and actual performance of the agent by the human in terms of the agent's abilities [YHSA20, MS06]. Much of previous research has looked at 'ability' as the core factor of estimating trust [YHSA20, EJS17, HBAD18] - *i.e. focusing upon the engineering aspect of trust.* However, as seen before, we propose to view trustworthiness as more than just ability. Our interpretation of trustworthiness can be enhanced when we not only focus upon agent capabilities but also on understanding integrity and benevolence which are often overlooked.

We would like to contrast our notions with the research that merely "promotes trust". We use the definition for developing appropriate trust from Hoffman et al.: "*A thorough understanding of both the psychological and engineering aspects of trust is necessary to develop an appropriate trust model*" [Hof17]. We identify a gap in the literature which focuses upon modelling integrity and benevolence of an artificial agent towards a human [URO13a]. In the following section 4.1.2, we propose a first attempt on how integrity can be modelled and in section 4.1.3 we refer our readers to one and only existing model for computational benevolence [URO13a] as per our knowledge *i.e. - focusing upon psychological aspects of trust.*

4.1.1 *Ab* - Ability

We can infer an agent's ability from the aggregation of all functionalities and capabilities it has. Understanding the agent's capabilities embedded by the developer can help both the agent and the human understand whether an agent has very low ability or very high ability. For this purpose, several existing trust-related functionality aggregators which focus on the system capabilities may be used, such as the ones described in [BJTT07, URO09, CLS+18]. In these functionality aggregators the trust is based on the task success, number of errors, skills and knowledge to accomplish a task. For example, competence beliefs about an agent [CF10], automation capability [KBDJ18] and agent's core functionality [IA19] are prominent inclusions for functionality aggregators. Therefore, in this paper we propose to rely on such existing aggregators to understand an agent's ability in terms of how it can form belief about its own trustworthiness.

4.1.2 *Ben* - Benevolence

Mayer et al. proposed that the effect of integrity on trust will be most salient early in the relationship prior to the development of meaningful benevolence [MDS95]. Therefore, we believe it is firstly important to understand how integrity can be understood and modeled as effect of perceived benevolence on trust will increase over time as the relationship between the parties develops. Moreover, as benevolence is about interpersonal relationships, it might not develop in agent-human relationships in the way it does for human-human ones. There are a number of steps taken in the social science research community to understand benevolence [LLS07, JAK+18]. However, we could only find one example of modelling benevolence in computer science community by Urbano et al. who classify benevolence as a Social Tuner in Human-AI interaction [URO13a]. According to them, Social Tuner measures the trustee's specific attachment toward the truster. This attachment is captured by the coefficient of benevolent actions. They estimate the value of the benevolence of the trustee toward the truster, $ben_{x,y}$, from the coefficient of benevolent actions. The coefficient of benevolent actions $\rho_{ba} \in [0, 1]$ measures the trend of contingencies presented by the trustee to the truster in the past.

$$ben_{x,y} = \frac{1}{2}\rho_{ba} + \frac{1}{2}\frac{cumValAgreem}{n} \tag{9}$$

where ρ_{ba} is the result of the correlation between the number of agreements established between truster and trustee in the past and cumulative value of past agreements ($cumValAgreem$). It is worth noting that the estimation of benevolence is only possible when there are, at least, two past interactions between the truster and the trustee under evaluation. The manner in which [URO13b] studies benevolence can also fit in our formalism focusing upon beliefs which forms specific relationship between trustee and trustor. Also, the value of benevolence must be updated at every new trustworthiness estimation, as benevolence may evolve due to the mutual (dis)satisfaction of the trustee with the relationship, which may change with time and context.

4.1.3 *I* - Integrity

We assume that an agent's integrity for a specific task is its integrity towards a human in accomplishing that specific task. We define integrity as the similarity of the human and agent values[1], meaning, having similar priorities over those values (which can be related to the actual definition of integrity focusing upon principles). In this section, we propose a basis for formally defining how integrity beliefs are formed. In particular, we derive two cases where an agent either possesses information regarding the values of the other or does not.

Case 1 We start by defining this relationship in the case where an agent has some belief about the values of the human. We state that the belief an agent a has about the integrity of a human h directly follows from the how similar agent a beliefs the priority ranking of their values to be to that of h:

$$if\ \{\mathcal{B}_a^{sim}(\mathcal{B}_a(priority_h(V_h)), priority_a(V_a))\} \uparrow\ then\ \{\mathcal{B}_a(I_a(h))\} \uparrow \tag{10}$$

$$if\ \{\mathcal{B}_a^{sim}(\mathcal{B}_a(priority_h(V_h)), priority_a(V_a))\} \downarrow\ then\ \{\mathcal{B}_a(I_a(h))\} \downarrow \tag{11}$$

where $\mathcal{B}_a^{sim}(X, Y)$ represents the belief of a about the similarity of X and Y, V_a is the value set of a, $priority_a(V)$ represents a priority ranking of agent a over this value set, and therefore $\mathcal{B}_a(priority_h(V_h))$ is a's belief about h's value and priority thereof. In other words, integrity beliefs of a about h are formed by a comparing their belief about h with what they know about themselves. We stipulate that if the belief about the similarity is higher, so will the belief in the integrity and *vice-versa*.

Case 2: In general it is important for an agent to rely on knowledge about human values. However, a situation could arise in which an agent has no information regarding a certain human teammate. In such a case we focus on the integrity *reputation* of the agent as per [ZBLA14]. We state that the integrity reputation of an cognitive agent (artificial or human) a_0 according to an artificial agent a, $IR_a(a_0)$, is the average sum of beliefs about a_0's integrity that are communicated (CB_a) by other autonomous agents $(a', a'', ...)$ to a. The equation 12 represents the communicated beliefs of a_0 according to a as sum of the communicated beliefs by other autonomous agents. Finally, the equation 13 represents the average sum of those beliefs forming the integrity

[1]Values are abstract motivations that justify opinions and actions, and are intrinsically linked to moral judgement [Sch12].

reputation of a_0 according to a.

$$CB_a(a_0, \{a', a''...\}) = \sum_{i=a'}^{a^n} \mathcal{B}_{a'}(I_{a'}(a_0)),\ \mathcal{B}_{a''}(I_{a''}(a_0))...\mathcal{B}_{a^n}(I_{a^n}(a_0)) \tag{12}$$

$$IR_a(a_0) = \frac{CB_a}{n} \tag{13}$$

5 Discussion

In this paper, we present a formalization of mutual appropriate trust. Our formalization allows us to structure the beliefs on trust and their dependencies in a human-agent dyadic relationship. This is crucial for understanding how we can promote appropriate mutual trust in human-agent teams. In this section, we will discuss the contribution of our formalization in the context of the previously presented dual-mode vehicle example (as in [MdS20]), which can both be driven autonomously by an artificial agent or by a human. We imagine that the default setting is the human driving according to the agent's instructions, but that the agent takes over when it recognizes dangerous situations. The human and the agent share the main goal of taking the human to their destination in a safe and pleasant way, while complying with societal ethics. As such, both the human and the vehicle have their set of values (e.g. their set of priorities when facing a hazardous scenario).

Agent's trust in the human

Using our formalization, we know trust is a directed belief on trustworthiness. That is, for an agent to trust a human, $T(a, h)$, to drive a vehicle safely and/or follow the agent's instructions, we first need to understand human's trustworthiness, $\mathcal{TW}_h(a)$, (e.g., whether the human can and will drive it safely in that environment, whether the human will follow a certain instruction). We then need to know how the beliefs of the agent $\mathcal{B}_a(\mathcal{TW}_h(a))$ are being formed, e.g., whether they are estimating ability (e.g., based on how the human is speeding, braking, keeping distance from other vehicles, etc), willingness (e.g., based on how many times the human ignores an instruction after seeing it on the panel, gets distracted, etc) or, perhaps, relying on other sources of information, such as reputation (e.g., reported accidents, whether the human followed other agents' instructions, etc). The agent's belief in human's trustworthiness should ideally correspond to the human's actual trustworthiness, i.e. $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h))) \equiv \mathcal{B}_h(\mathcal{TW}_a(h))$, meaning the agent appropriately trusts the human.

Human's trust in the agent

It is also important that the agent is able to estimate a human's trust towards the agent itself, $T(h, a)$, both when estimating human's trustworthiness (e.g. knowing whether the human will trust and follow agent's instructions may facilitate the prediction of human's actions), and for appropriate trust elicitation (e.g. if the agent detects the human is not trusting agent's actions, it can alter its behaviour). As such, we need to understand how a human trusts an agent and how a belief about this trust is formed, i.e. $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h)))$. In this paper we propose that trustworthiness, as well as trust, are directed, and illustrate this using the *ABI* model. This provided us a tool to model human's trust from both engineering lens, *Ability* (e.g. capability of driving the vehicle and following the instructions), and psychological lens, *Benevolence & Integrity*, which depends on the agent too (e.g. complying with agent's instruction upon hazardous scenarios). In particular, we define integrity as the similarity between human's and agent's values. This means that by knowing humans' values (e.g. helping others is a priority), the agent can estimate human's trust in itself according to their value similarity (e.g. the human may trust it more if the agent stops when it detects an accident, calls emergency services, and asks human to check on the possible victims). The agent's belief in the human's trust towards the agent should correspond to the human's actual trust in the agent, i.e. $\mathcal{B}_a(\mathcal{B}_h(\mathcal{TW}_a(h))) \equiv T(h, a)$.

Appropriate trust elicitation through nested beliefs

The inverse estimate of human trust (*i.e.* from an agent towards a human) allows an agent to form a belief in its own performance of the driving task/giving instructions, and use that information to estimate the corresponding influence of its trustworthiness (i.e., increasing, decreasing, constant) [JFC+18]. Following the example of integrity from the previous paragraph, by knowing humans' values (e.g. helping others is a priority), the agent could for instance adjust its values accordingly and possibly increase its trustworthiness (e.g. stopping when it

detects an accident and calling emergency services). Consequently, if the human's trust in the agent increases, the agent's trust in the human may increase too (e.g. the agent now knows the human will follow the instructions that comply with human's values). For this, agent's belief on its own trustworthiness should correspond to its actual trustworthiness, i.e. $B_a(\mathcal{TW}_a(h)) \equiv \mathcal{TW}_a(h)$. By guaranteeing both human's and agent's appropriate trust, i.e. their mutual trust corresponds to their mutual trustworthiness, we have appropriate mutual trust. The understanding that these definitions are nested and that they depend on each other, provides support for the design of human-agent interaction.

6 Limitations & Future Work

To appreciate the contribution of this work, we also need to understand its limitations. In our definitions for trustworthiness and trust in agents or humans, we rely on the belief of one in another. However, we explicitly do not focus on how those beliefs are generated and how we can evaluate them. Although these are crucial for achieving mutual appropriate trust, we argue that we can only start to generate and evaluate beliefs once we understand what they are about. This work takes this first crucial step. Moreover, we note that according to many models, including the trust model by Rahman & Hailes [ARH00], trust beliefs will involve beliefs of risk, utility and motivations along with trustworthiness. This means that a study into how beliefs about trustworthiness are formed should also take these aspects into account. Finally, it is worthwhile quoting Nilsson [Nil14], who warns researchers about "belief traps" *i.e.* holding onto beliefs that wouldn't survive critical evaluation. Nilsson states it is important to expose beliefs to the reasoned criticism of others. Yet, we do not explicitly incorporate other agents criticism for the formation of beliefs regarding trust or trustworthiness. However, we believe this can be formally extended in future work when, for example, agent a_1 is forming the belief regarding human h_1's trust in itself, a_1 can take into account another agent's criticism (if this other agent has interacted with h_1 before).

Additionally, in section 3, expressions 3, 4 and 5 express the ultimate goal of this proposition of formalization. That is, according to expression 3, how the agent's trust in a human should ultimately be equivalent to the actual human's trustworthiness. Similarly, according to Expression 4, how agent's belief on human's trust should be ultimately equivalent to 1) human's trust and 2) to agent's trustworthiness towards the human. Finally, Expression 5 suggests that the agent's belief in its own trustworthiness should correspond to its actual trustworthiness. We do not offer a thorough evaluation of these expressions at this point. However, we feel that the concept of nested beliefs is already a relevant contribution since it helps us to understand what needs to be taken into account when designing a system dependent on appropriate mutual trust. Going forward, we aim to refine the formalization and implement a method to evaluate it. In particular, we would like to conduct user studies to evaluate our notions regarding beliefs of beliefs in an experimental setting. This can both help us understand how trust beliefs are formed in humans, and how agents can appropriately use these beliefs to actually improve teamwork. Another interesting direction for future research would be to study if the trust of a human h in an agent a_1 is affected by the trust of h in an agent a_2 or the trust of a_1 in a_2.

7 Conclusion

In this paper we presented a formalization of appropriate mutual trust in a dyadic human-agent relationship in the context of teamwork. We discussed how forming beliefs in both trust and trustworthiness (of human teammates and agent's own) is crucial for any agent in a human-agent team. Particularly, we formalized trust as a belief of directed mutual trustworthiness. This belief should correspond to the actual trustworthiness, supporting appropriate mutual trust. We looked at the specific beliefs in trust and trustworthiness that affect 1) agent's appropriate trust in a human teammate and 2) human's appropriate trust in an agent teammate, how these beliefs are nested, and illustrated these beliefs with the *ABI* model. All in all, if an agent is able to reason about trust, trustworthiness, and their formation, it will also be able to use appropriate trust to make better choices of actions, such as task delegation and risk mitigation. Moreover, the agent can also calibrate its own trustworthiness to elicit human's appropriate trust, leading to a proper use of technology, more effective teamwork, and more safety for the human teammates.

Acknowledgments

This work is part of the research lab AI*MAN of Delft University of Technology.

References

[ARH00] Alfarez Abdul-Rahman and Stephen Hailes. Supporting trust in virtual communities. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*. IEEE, 2000.

[ASKL19] Ighoyota Ben Ajenaghughrure, Sonia C Sousa, Ilkka Johannes Kosunen, and David Lamas. Predictive model to assess user trust: a psycho-physiological approach. In *Proceedings of the 10th Indian Conference on Human-Computer Interaction*, pages 1–10, 2019.

[BHHH20] Christina Breuer, Joachim Hüffmeier, Frederike Hibben, and G. Hertel. Trust in teams: A taxonomy of perceived trustworthiness factors and risk-taking behaviors in face-to-face and virtual teams. *Human Relations*, 73:3 – 34, 2020.

[BJTT07] Tibor Bosse, Catholijn M Jonker, Jan Treur, and Dmytro Tykhonov. Formal analysis of trust dynamics in human and software agent experiments. In *International Workshop on Cooperative Information Agents*, pages 343–359. Springer, 2007.

[BNS13] Chris Burnett, Timothy J. Norman, and Katia Sycara. Stereotypical trust and bias in dynamic multiagent systems. *ACM Transactions on Intelligent Systems and Technology*, 4, 3 2013.

[CF00] Cristiano Castelfranchi and Rino Falcone. Trust is much more than subjective probability: Mental components and sources of trust. In *Proceedings of the 33rd annual Hawaii international conference on system sciences*. IEEE, 2000.

[CF10] Cristiano Castelfranchi and Rino Falcone. *Trust Self-Organising Socio-technical Systems*. Springer International Publishing, 2010.

[CLS+18] Jessie YC Chen, Shan G Lakhmani, Kimberly Stowers, Anthony R Selkowitz, Julia L Wright, and Michael Barnes. Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science*, 19(3):259–282, 2018.

[CMP17] Caterina Cruciani, Anna Moretti, and Paolo Pellizzari. Dynamic patterns in similarity-based cooperation: An agent-based investigation. *Journal of Economic Interaction and Coordination*, 12(1), 2017.

[CNGD19] Kinzang Chhogyal, Abhaya C. Nayak, A. Ghose, and Khanh Hoa Dam. A value-based trust assessment model for multi-agent systems. *28th International Joint Conference on Artificial Intelligence (IJCAI-19)*, 2019.

[EJS17] Fredrick Ekman, Mikael Johansson, and Jana Sochor. Creating appropriate trust in automated vehicle systems: A framework for hmi design. *IEEE Transactions on Human-Machine Systems*, 48(1):95–101, 2017.

[FDA14] Michael W Floyd, Michael Drinkwater, and David W Aha. How much do you trust me? learning a case-based model of inverse trust. In *International Conference on Case-Based Reasoning*, 2014.

[FDA16] Michael W Floyd, Michael Drinkwater, and David W Aha. Learning trustworthy behaviors using an inverse trust metric. In *Robust Intelligence and Trust in Autonomous Systems*. Springer, 2016.

[FPVC13] Rino Falcone, Michele Piunti, Matteo Venanzi, and Cristiano Castelfranchi. From manifesta to krypta: The relevance of categories for trusting others. *ACM Transactions on Intelligent Systems and Technology*, 4, 3 2013.

[Gri05] N. Griffiths. Task delegation using experience-based multi-dimensional trust. In *AAMAS '05*, 2005.

[GY20] Yaohui Guo and X. Jessie Yang. Modeling and predicting trust dynamics in human–robot teaming: A bayesian inference approach. *International Journal of Social Robotics*, 2020.

[HBAD18] Sandy H Huang, Kush Bhatia, Pieter Abbeel, and Anca D Dragan. Establishing appropriate trust via critical states. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3929–3936. IEEE, 2018.

[HJW19] H. Huynh, C. E. Johnson, and Hillary S Wehe. Humble coaches and their influence on players and teams: The mediating role of affect-based (but not cognition-based) trust. *Psychological Reports*, 123:1297 – 1315, 2019.

[HLHV09] Andreas Herzig, Emiliano Lorini, Jomi F. Hübner, and Laurent Vercouter. A logic of trust and reputation. *Logic Journal of the IGPL*, 18:214–244, 12 2009.

[Hof17] Robert R Hoffman. A taxonomy of emergent trusting in the human–machine relationship. *Cognitive systems engineering: The future for a changing world*, pages 137–163, 2017.

[IA19] Brett W Israelsen and Nisar R Ahmed. Dave... i can assure you... that it's going to be all right... a definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships. *ACM Computing Surveys (CSUR)*, 51(6):1–37, 2019.

[JAK+18] Theodore Jensen, Yusuf Albayram, Mohammad M.H. Khan, Ross Buck, Emil Coman, and Md A. Al Fahim. Initial trustworthiness perceptions of a drone system based on performance and process information. In *Proceedings of 6th International Conference on Human-Agent Interaction*, 2018.

[JFC+18] Benjamin Johnson, Michael W Floyd, Alexandra Coman, Mark A Wilson, and David W Aha. Goal reasoning and trusted autonomy. In *Foundations of trusted autonomy*. Springer, Cham, 2018.

[KBDJ18] Siddartha Khastgir, Stewart Birrell, Gunwant Dhadyalla, and Paul Jennings. Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles. *Transportation research part C: emerging technologies*, 96:290–303, 2018.

[LLS07] Dong-Jin Lee, Moonkyu Lee, and Jaebeom Suh. Benevolence in the importer-exporter relationship. *International Marketing Review*, 2007.

[LLS20] Michael Lewis, Huao Li, and Katia Sycara. Deep learning, transparency, and trust in human robot teamwork. In *Trust in Human-Robot Interaction*, pages 321–352. Elsevier, 2020.

[LS04] J. D. Lee and Katrina A. See. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 46:50 – 80, 2004.

[LSW18] Michael Lewis, Katia Sycara, and Phillip Walker. *The Role of Trust in Human-Robot Interaction*. Springer International Publishing, 2018.

[MDS95] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Source: The Academy of Management Review*, 20:709–734, 1995.

[MdS20] Giulio Mecacci and Filippo Santoni de Sio. Meaningful human control as reason-responsiveness: the case of dual-mode vehicles. *Ethics Inf. Technol.*, 22(2):103–115, 2020.

[MS06] John M McGuirl and Nadine B Sarter. Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. *Human factors*, 48(4):656–665, 2006.

[NGP+20] Catherine Neubauer, Gregory Gremillion, Brandon S. Perelman, Claire La Fleur, Jason S. Metcalfe, and Kristin E. Schaefer. Analysis of facial expressions explain affective state and trust-based decisions during interaction with autonomy. In *Proceedings of the 3rd International Conference on Integrating People and Intelligent Systems, February 19-21, 2020, Modena, Italy*, volume 1131 of *Advances in Intelligent Systems and Computing*, pages 999–1006. Springer, 2020.

[Nil14] Nils J Nilsson. *Understanding beliefs*. MIT Press, 2014.

[NPW18] Andrew M. Naber, Stephanie C. Payne, and Sheila Simsarian Webber. The relative influence of trustor and trustee individual differences on peer assessments of trust. *Personality and Individual Differences*, 128:62–68, 7 2018.

[NWL+20] Changjoo Nam, Phillip Walker, Huao Li, Michael Lewis, and Katia Sycara. Models of trust in human control of swarms with varied levels of autonomy. *IEEE Transactions on Human-Machine Systems*, 50:194–204, 6 2020.

[OSPJ13] Scott Ososky, David Schuster, Elizabeth Phillips, and F. Jentsch. Building appropriate trust in human-robot teams. In *AAAI Spring Symposium: Trust and Autonomous Systems*, 2013.

[RG95] Anand Srinivasa Rao and M. Georgeff. Bdi agents: From theory to practice. In *ICMAS*, 1995.

[Sch12] Shalom H Schwartz. An overview of the schwartz theory of basic values. *Online readings in Psychology and Culture*, 2(1):2307–0919, 2012.

[SMD07] F. Schoorman, Roger Mayer, and J. Davis. An integrative model of organizational trust: Past, present, and future. *Academy of Management Review*, 32:344–354, 2007.

[SMV13] Jordi Sabater-Mir and Laurent Vercouter. Trust and reputation in multiagent systems. *Multiagent systems*, page 381, 2013.

[SPG+21] Kristin E. Schaefer, Brandon S. Perelman, Gregory M. Gremillion, Amar R. Marathe, and Jason S. Metcalfe. A roadmap for developing team trust metrics for human-autonomy teams. In *Trust in Human-Robot Interaction*. Academic Press, 2021.

[SSB05] E. Salas, Dana E. Sims, and C. Burke. Is there a "big five" in teamwork? *Small Group Research*, 36:555 – 599, 2005.

[SW19] Vidullan Surendran and A. Wagner. Your robot is watching: Using surface cues to evaluate the trustworthiness of human actions. *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–8, 2019.

[UG20] Anna-Sophie Ulfert and Eleni Georganta. A model of team trust in human-agent teams. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 171–176, New York, NY, USA, 2020. Association for Computing Machinery.

[URO09] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. Computing confidence values: Does trust dynamics matter? In *Portuguese Conference on Artificial Intelligence*, pages 520–531. Springer, 2009.

[URO13a] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. The impact of benevolence in computational trust. In *Agreement Technologies*, pages 210–224. Springer, 2013.

[URO13b] Joana Urbano, Ana Paula Rocha, and Eugénio Oliveira. A socio-cognitive perspective of trust. In *Agreement Technologies*, pages 419–429. Springer, 2013.

[VMP+20] Ewart J De Visser, Marieke, M M Peeters, Malte, F Jung, Spencer Kohn, Tyler, H Shaw, Richard Pak, and Mark A Neerincx. Towards a theory of longitudinal trust calibration in human-robot teams. *International Journal of Social Robotics*, 12:459–478, 2020.

[VPCC19] Samuele Vinanzi, Massimiliano Patacchiola, Antonio Chella, and Angelo Cangelosi. Would a robot trust you? developmental robotics model of trust and theory of mind. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374, 4 2019.

[WA11] Alan R Wagner and Ronald C Arkin. Recognizing situations that demand trust. In *2011 RO-MAN*, pages 7–14. IEEE, 2011.

[Win18] Michael Winikoff. Towards trusting autonomous systems. *Lecture Notes in Computer Science*, 10738 LNAI:3–20, 2018.

[WRH18] Alan R. Wagner, Paul Robinette, and Ayanna Howard. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems*, 8, 11 2018.

[YHSA20] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. How do visual explanations foster end users' appropriate trust in machine learning? In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 189–201, 2020.

[ZBLA14] Yuhui Zhong, Bharat Bhargava, Yi Lu, and Pelin Angin. A computational dynamic trust model for user authorization. *IEEE Transactions on Dependable and Secure Computing*, 12(1):1–15, 2014.