automation

MDPI

Article

Modeling Interaction in Human–Machine Systems: A Trust and Trustworthiness Approach

Alessandro Sapienza *, Filippo Cantucci and Rino Falcone

Institute of Cognitive Sciences and Technologies, National Research Council of Italy (ISTC-CNR), 00185 Rome, Italy; filippo.cantucci@istc.cnr.it (F.C.); rino.falcone@istc.cnr.it (R.F.)
* Correspondence: alessandro.sapienza@istc.cnr.it

Abstract: Trust has been clearly identified as a key concept for human–machine interaction (HMI): on the one hand, users should trust artificial systems; on the other hand, devices must be able to estimate both how much other agents trust them and how trustworthy the other agents are. Indeed, the applications of trust in these scenarios are so complex that often, the interaction models consider only a part of the possible interactions and not the system in its entirety. On the contrary, in this work, we made the effort to consider the different types of interaction together, showing the advantages of this approach and the problems it allows to face. After the theoretical formalization, we introduce an agent simulation to show the functioning of the proposed model. The results of this work provide interesting insights for the evolution of HMI models.

Keywords: trust; trustworthiness; cognitive modeling; intelligent systems; artificial intelligence; HMI

Citation: Sapienza, A.; Cantucci, F.; Falcone, R. Modeling Interaction in Human–Machine Systems: A Trust and Trustworthiness Approach. *Automation* **2022**, *3*, 242–257. https://doi.org/10.3390/automation3020012

Academic Editor: Alessandro Gardi

Received: 10 February 2022
Accepted: 22 March 2022
Published: 25 March 2022

1. Introduction

Over the last years, trust has been identified as a key feature in human–AI interaction. In scenarios such as multi-agent systems (MAS), Internet of Things (IoT), and human–robot interaction (HRI), humans need to trust the artificial systems they need to collaborate with [1,2]. Indeed, current trends suggest a growing presence of autonomous artificial entities in human environments. Intelligent objects, robots, personal assistants, etc. will play an increasingly important role in our lives, integrating into our daily activities and carrying out elaborated tasks for us [3,4].

In the light of these premises, trust represents a fundamental tool to handle the complex relationships arising [5,6]. On the one hand, this means that artificial systems should possess and expose a wide range of characteristics that define their trustworthiness. However, this is not the only type of relationship that occurs in these contexts. Being such a pivotal topic, many works have contributed to the creation of trust models in the field of human–robot or human–AI interaction. Regrettably, most of the existing approaches focus just on a specific dimension of the trust relationship (trust or trustworthiness), thus not analyzing the interaction in its entirety. Although this approach represents a necessary step, it also introduces a series of limitations and does not allow to fully exploit the potential of artificial systems. In order to understand well the current limitations of the state of the art, it is first necessary to define properly the different types of trust relationships existing in this domain.

Figure 1 provides a clarification of the link between trustworthiness and trust. Trustworthiness is an intrinsic property of the trustee, which accounts for his actual ability and willingness to carry out the tasks. Since this is an internal characteristic of the trustee, it cannot be accessed directly but only estimated. Therefore, to evaluate a trustee, a trustor must have a trustworthiness evaluation model. Trustworthiness evaluation is generally used for the purpose of selecting reliable partners for interaction. In addition to this, within this framework, the trustee can reason about the trustor's perception of it: how trustworthy the trustor thinks it is. Therefore, this time, the trustee realizes a model of the trustor's

mind in order to assess how the trustor trusts it. In that case, we will talk about trust evaluation. These models are generally used to allow trustees to understand how much they are trusted and how they should behave to be perceived more trustworthy. Regarding the relationship between trust and trustworthiness, it is important to underline that this relationship is indeed reciprocal and not one-way. On the one hand, the trustor modifies its trust based on the trustworthiness of the trustee. However, the trustee can also adjust its trustworthiness due to the trustor's trust. For example, if the trustee believes that the trustor has a too low level of trust, it may act by trying to determine the conditions under which it is more trustworthy. In other words, it is a two-way relationship, in which trustor and trustee can influence each other.

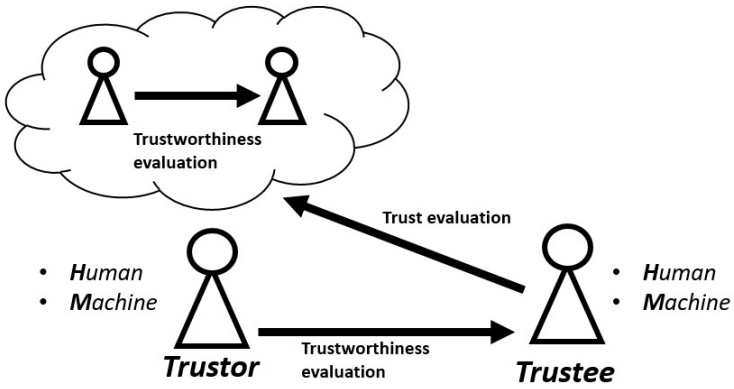


Figure 1. Relationship between trustworthiness and trust.

As a further limitation, trust models often limit the role of trustor and trustee specifically to humans or artificial systems, i.e., they rarely consider they possibility of hybrid systems allowing human and artificial agents to interchangeably cover these roles.

Remarkably, how and to which extent humans should trust such artificial systems has been strongly investigated and has attracted particular interest (human–machine models, H2M). Consequently, the current literature in this area has mainly addressed the issue of the trustworthiness of these systems (What characteristics should an artificial system have to be trustworthy?). For instance, Refs. [7,8] propose two computational models, respectively, for IoT and MAS where users evaluate the trustworthiness of the artificial devices based on direct satisfaction experiences, past interaction experiences, and recommendations from others. As a further example, Zhong et al. [9] propose a dynamic trust model for user authorization, based on the direct experience, recommendation, competence, and integrity of the users involved in the multi-agent system.

In a similar way, in machine–machine (M2M) models, devices interact with each other to perform tasks for their users. To this purpose, M2M models focus on trustworthiness [10,11], since the artificial systems need to evaluate the trustworthiness of their peers, based on different properties, regarding both the artificial system itself and the human who owns it. In the domain of IoT, Ba-hutair et al. [12] introduced a multi-perspective trust model to assess the trust value of a crowdsourced IoT service. Fortino et al. [13] consider a model based on reliability, local reputation, distance, and helpfulness. Wang et al. [14] proposed a mechanism to evaluate nodes' trustworthiness from data collection and communication behavior.

These models tend to be very elaborate on the interaction between devices, while the user has a marginal role or no role at all. In some cases, user preferences are taken into account. However, the human–machine interaction is neglected.

As far as trust is concerned, human–machine models (H2M) speculate a lot on how devices should act in order to be (and to be perceived) as more trustworthy as possible, with respect to their user. Especially in HRI scenarios, different solutions are provided [15,16], in order to build robots able to evaluate the trust that their human interlocutors have in them. Among the many, Edmonds et al. [17] investigate the effect of explainability, Nikolaidis et al. [18] consider a model of mutual adaptation. Chen et al. [19] propose a model where robots

can even intentionally fail (having no other communication options) in order to correct the overestimation of its actual ability and to adjust the trust evaluation of its interlocutor. Regarding the same domain, Cantucci and Falcone [20] propose a cognitive architecture for trustworthy human–robot collaboration. This is accomplished using Theory of Mind (ToM), which is the psychological ability to assign to others beliefs and intentions that can differ from one's owns. A problem of H2M trust models is that while the human–machine relationship is very articulated, it is often limited to a one-to-one relationship. While this may be adequate for many studies and applications, we are instead interested in situations in which users have multiple devices available; on the other hand, devices may have to manage multiple users. In addition, a one-to-one model does not allow devices to deal with situations they cannot manage on their own (errors, lack of resource, etc.), since such models do not provide the necessary tools to let devices decide how to behave in these cases. In both cases, the role of the machine is limited precisely because a limited model is considered.

In the cases examined above, machines assess human trust toward themselves. Although this is the predominant topic, other works have also begun to consider the trustworthiness of artificial agents toward humans (M2H), namely the situation in which a device acts as a trustor requiring the execution of a task to a human trustee [21]. This type of interaction is increasingly common when considering collaborative scenarios: the possibility to estimate the ability of other agents in performing tasks allows the devices to understand if and when it is better to do tasks on their own or delegate them to a user [22]. This ability is essential if we aim to provide a real hybrid environment in which humans and artificial agents collaborate to achieve common goals. Such a problem has been mainly investigated in the HRI domain. For example, Vinanzi et al. [23] propose an artificial cognitive architecture based on the developmental robotics paradigm that can estimate the trustworthiness of the human actors for the purpose of decision making.

Remarkably, the complexity of the interaction led also to the development of computational models giving artificial agents the possibility to estimate how much other agents trust them [24], whether they are other artificial entities or human users.

With the exception of multi-agent systems, proposing even approaches [25,26] that generalize the trustworthiness evaluation of agents against other agents, regardless of whether they are humans or machines, all of the above approaches focus on the interaction between specific agents, human or artificial, and they do not propose solutions that overcome this kind of specification. Therefore, despite their high level of accuracy, such models are designed and built to fit a specific kind of interactions and consider only part of the complex relationship of trust.

On the one hand, introducing a complete model capable of considering all the trust relationships together allows us to exploit the information that one type of interaction provides in favor of the others, making the system more flexible to possible situations that may arise. On the other hand, we need to start rethinking the role of devices. These can have an effective active role in the system precisely because they are at the center of the interaction.

In order to cope with such a research gap, in this work, we try to address the issue of trusted collaboration between agents (be they human or artificial) in its most complete application: trustor and trustee can be interchangeably an artificial or a human agent. Therefore, we will not simply address the problem of artificial systems' trustworthiness but also that of their trust (What characteristics must an artificial system have to trust another system?) and their reciprocal relationships.

Within this system, albeit autonomously, the artificial agents perform tasks to achieve the goals of their human users. So, even when they are the promoters of new tasks, the ultimate goal is always to satisfy the necessity of the community to which their partners belong. Yet, precisely for this reason, it is possible that the actions of the autonomous devices may in some cases conflict with those of the single users, even in the attempt to offer the best possible performance. This is why such artificial entities cannot be simple

executors of tasks. Rather, they represent complex entities capable of making autonomous decisions, not necessarily corresponding to what is directly asked of them.

After providing a theoretical formulation of such systems, we introduce a simulation to study how it is possible to concretely use this framework with respect to both trustworthiness and trust.

The contributions of this work are as follows:

1. We propose an innovative trust approach for HMI, aiming to provide a complete vision, thus considering humans and artificial agents both from the trustor's and from the trustee's side;
2. We discuss the added values introduced by this approach, highlighting the way it allows us to overcome some problems and model aspects and phenomena that were previously missing;
3. Considering trust in AI systems, it is important to underline that while they need to understand how humans trust them, they should also aim to promote proper trust [27]. This means that an artificial device should guide its interlocutor to trust it according to its concrete potential performances. In fact, a system that wants to be truly collaborative and that aims to support the goals of its interlocutor should not aim to be perceived reliable beyond its abilities, otherwise it would be deceptive, but rather it should provide the correct perception of its abilities.

It is worth underlining that the purpose of this contribution is not to quantify the actual benefits introduced by the proposed modeling, as this would be strongly domain-specific and would require a field study. Rather, we intend to show the actual potential of our approach, the limits it allows to overcome, and the added value it brings to this type of system. This section has highlighted the research gaps and the contribution we intend to realize. In addition, the rest of the article is organized as follows. In Section 2, we introduce the theoretical formulation of our work, together with the related concepts and issues. On the basis of these theoretical premises, Section 3 discusses the implementation we realized, whose results are presented in Section 4 and discussed in Section 5. In conclusion, Section 6 summarizes the contribution of the whole work.

2. Theoretical Formulation

In this framework, we are interested in producing a general formulation of the complex relations between humans $\mathcal{H}$ and artificial entities $\mathcal{E}$ within hybrid environments. We will generally refer to them as agents $\mathcal{A}$. When a given agent a_i needs to carry out a specific task τ that it cannot or does not want to execute itself, a_i needs to find a reliable partner to carry out τ for it. Of course, the choice of the partner can be made randomly. Nevertheless, a random choice exposes the trustor to several risks: it may delegate a task to an incompetent, unwilling, or even malicious partner. In these cases, trust assessment has a clear and precise role, because it allows to reduce the risk of being exposed to such problems. Thus, rather than choosing randomly, a_i ranks all the possible partners according to their trustworthiness. Therefore, the trustor selects those partners who would provide sufficient performance, i.e., having a trustworthiness value above a certain internal threshold σ, and it goes through the ranked list until one of the agents accepts τ. Since we are interested in modeling a hybrid system, devices may require the execution of tasks to other devices or even to human partners. Thus, we considered three kinds of interaction: human to device (H2M), device to device (M2M), and device to human (M2H). The structure of these three types of requests is similar. Of course, the behavior of the trustee changes in relation to the type of trustor.

On the other side, upon the occurrence of a task request τ, an agent a_j will consider whether to accept the task or not. Indeed, a_j has its own goals and priorities, and it will decide according to them. In our case, we assume that the main priority of the agents is to satisfy the requests of the other agents in the system in such a way as to optimize the overall performance of the system itself. Therefore, if a_j is free, it will always accept τ. If it is not, it must consider different strategies to choose how to behave. First of all, they may

check whether it is possible to *postpone the task*. A trustee first checks if the new task can be postponed, i.e., if the trustor accepts that the a_j executes τ after the task it is currently doing. If this is not the case, in a similar way, a_j tries to postpone the old task. When this is not possible, one of the two tasks can be *reassigned* to someone else. At this point, the trustee tries to relocate one of the two tasks to another device. It is worth noting that this process is not the equivalent of letting the trustor handle the task itself. In fact, if the trustee manages the process, it has the possibility to introduce an optimization mechanism. Given that the trustee knows both its and the other agents' ability to perform both the old and the new tasks, it will try to allocate tasks in order to best fit trustees' actual capabilities. In the end, if neither task can be postponed nor reassigned, one of the tasks must be aborted. A simpler strategy is to reject the new task. We suppose that a priority parameter can be assigned to tasks. Then, the trustee may decide to carry out the task with the highest probability and to interrupt/refuse the other one. Of course, rejecting the new task is the same as saying that it has lower priority than the old one.

2.1. Trustworthiness Evaluation

Evaluating the trustworthiness of the potential partners is essential to understand who to interact with. For trust modeling, we were inspired by the socio-cognitive model of trust developed by Castelfranchi and Falcone [28]. Specifically, in accordance with Catelfranchi and Falcone, trust assessment is based on two specific dimensions:

- *Competence* represents the set of qualities making the trustee good for the task τ: skills, know how, expertise, knowledge, self-esteem, self-confidence, and so on.
- *Willingness* is a prediction of the trustee's behavior: the fact that it is reliable, predictable, that we can count on it. Willingness evaluates whether the trustee is willing and persistent [29,30] (it is not prone to give up, but it will insist on doing τ). As we are considering a system designed to best meet user requests, we assume that devices are always willing to cooperate. However, given the presence of multiple trustors, the devices could be engaged in other requests, so their *availability* must be taken into account. Furthermore, since we are considering a physical context, *physical proximity* can impact on availability, if this is important for the realization of the task.

Different information can be used to produce a trust evaluation. The most immediate one is direct experience. In this case, the trustor a_i evaluates the trustee a_j based on direct past interactions. The clear advantage of this approach is the use of direct information; since there is no intermediary, no uncertainty is introduced due to the reliability of the information sources. Nevertheless, direct experience requires a certain number of interactions to produce a good evaluation, and initially, a_i should trust a_j without any reason to do so (the cold start problem). It is also possible to rely on second-hand information, exploiting recommendation [31] or reputation [32]. In this case, there is the advantage of having a ready-to-use evaluation. The disadvantage is that this evaluation introduces uncertainty due to the recommender's ability and benevolence. Lastly, it is possible to use inference and knowledge generalization, such as categories or stereotypes [25]. A category is a general set of agents—doctors, thieves, dogs, and so on—whose members have common characteristics, determining their behavior or their ability. If it is possible to associate a_j to a category C and the average performance of the members of C concerning τ is given, I can exploit this information to assess a_j's trustworthiness. The advantage is that it is possible to evaluate every node belonging to a known category, even if no one knows it. The disadvantage is that the level of uncertainty due to this method can be high, depending on the variability inside the category and its granularity.

Since in this work, we are interested in showing the practical use of our framework, we will focus on direct experience and categorization. Within our model, we assume that it is always possible to know the category to which the agents belong and that the performance of the category for the various tasks is given. We do not intend to go into detail on the use of categories, as we do not intend to study it here. We are only interested in showing their use, while their effectiveness has widely been demonstrated in the literature [26,33,34]. As

for recommendations, as we will see in Section 2.2, we will consider only a specific sub-case, which is of particular interest here.

2.2. Reasoning about Trust

A point of particular interest on which we want to focus is that of trust. Surely, assessing the reliability of partners is fundamental. However, reasoning on the trustee's side, analyzing trust, i.e., estimating trustors' perception of the trustee's reliability, introduces a strong added value. In many works, trust reasoning is introduced with the purpose of maximizing the perception that humans have of devices. On the contrary, we argue that the ultimate goal of the system should not be to maximize the perceived reliability of the user but rather to perform the tasks for the users in the best possible way, also making the user perceive the corresponding level of reliability. Paradoxically, trying to maximize user trust could even be counterproductive: it is important that users trust devices correspondingly to their and to the system's possibilities in terms of competence (actual device capacity) and availability (it depends system load, i.e., how many tasks are generated). Indeed, overestimating the trustee's ability is counterproductive because the trustor will base its decision on a wrong estimation not coinciding with the real possibilities of the trustees. Having more precise information, the trustor could have chosen better. Then, devices should aim to *proper trust*. In light of these considerations, estimating the perception that the trustor has of me represents an added value if I use such knowledge to help the trustor in its decisions. In our specific case, when the trustee finds out that the trustor believes it is less reliable than it thinks, the trustee can act in order to modify the trustor perception by being more available to the trustor, for instance, getting closer to it. By doing so, the trustee has a higher chance of being chosen and improving the trustor's perception. Furthermore, the trustee may discover that it is considered more reliable than it really is. Since, in general, devices perform the same task for different trustors, they have more accurate data on their own reliability. Therefore, when a trustor chooses them because it believes them to be much more reliable than reality, they can inform the trustor of their actual capabilities and propose possible alternatives.

2.3. Presence of Error

Beyond the actual ability of a device to perform a task, occasional errors may occur due to external conditions. For example, a device could be asked to bring a drug to a patient, but that drug may have run out. Furthermore, the device might find an insurmountable obstacle along its way. In both cases, the device is unable to complete the task because of an external situation, i.e., independent of its ability to perform the requested task. Note that an error is different from a bad performance. The last one implies that the trustee completes the task, but it does so in an unsatisfactory way. This relates to the trustee competence; thus, such a situation is interpreted as a bad performance by the trustor. On the contrary, when we talk about error, we mean a problem caused by an external source, dependent on the context (on which the trustee is not intended to intervene). If the trustee is able to identify the error and compensate by requiring human intervention, this event should treated differently by the trustor.

If we only look at the H2M relationship, in the presence of such a situation, the trustee has no choice but to let the task fail. On the other hand, if we consider the system in its entirety, the trustee can generate a new task in which it requires the intervention of another agent able to solve the problem introduced by the error. Assuming that the other devices, similarly to himself, do not have the required capabilities, the trustee can request the intervention of a human agent (M2H). To do this, of course, the device must also be able to evaluate human agents as executors of tasks. Thus, when the device detects the presence of an external error, it can request for assistance, generating a new subtask to solve the error. This represents an added value as the task is not left to fail. Of course, the final result depends on whether the user completes the subtask or not.

2.4. Devices as Mediator for Task Reassignment

Previously, we discussed the possibility for a device to reassign its tasks. Indeed, it may happen that while trustee a_j is executing τ_i for the trustor a_i, another agent a_k may require the execution of a new task τ_k. In this case, if we think of the device as a simple task executor, the only possible choices are refusing the new task or accepting it, interrupting the old one. However, if we consider the device as an active part of the system, endowed therefore with the ability to assign tasks to other devices (M2M), thus evaluating their performance, a_j can introduce optimization mechanisms. Indeed, since it knows both its and others' trustworthiness for the two tasks, it is able to reallocate the tasks in order to optimize the final result. This is possible only if we look at the system in its entirety and not simply as made of individual interactions. In fact, the following are necessary:

1. a_j is able to act as a trustor with respect to other devices, assigning tasks according to trustworthiness assessment;
2. The interaction model can consider more (kinds of) trustor–trustee relations at the same time.

Note that when a task τ_i, assigned by trustor a_i to trustee a_j, is reassigned to a third agent a_l, a_j acts as a trustor with respect to a_l; thus, both a_i and a_j evaluate a_l on its ability to perform τ_i. In other words, the role of mediator allows a_j to acquire further information about the other devices. In turn, this increases its ability as a mediator.

As far as it concerns the relationship between a_i and a_j after reassignment, since the latter has not carried out τ_i, its performance cannot be evaluated. Nevertheless, a_j could be evaluated by a_i for its ability as mediator. Thus, in the future, a_i may decide to trust a_j's suggestion or to handle reassignment itself according to such information.

3. Simulations

The model we introduced in the previous sections is deliberately abstract, in order to be as general as possible. Conversely, in the simulations, we considered a hypothetical application scenario, aiming to show how the model can be used in practice. Specifically, we considered the case of an intelligent assistance system within a hospital environment. In this case, the environment is populated by human agents as doctors, nurses, and patients $\mathcal{H} = \{D, N, P\}$ and healthcare robots with different features and capabilities $\mathcal{E} = \{E_1, E_2, E_3\}$.

Let us consider three kinds of task: τ_1 and τ_2 requiring physical presence (such as assistance or object transportation) and τ_3 not requiring it (such as remote assistance). In addition, there is a task specific for each category of human. For these tasks, devices act only as a trustor, as we suppose they are not able/authorized to execute them. Overall, a task is modeled in terms of

1. *Type*: τ_1, τ_2, τ_3, τ_{doctor}, τ_{nurse}, $\tau_{patient}$;
2. *Execution time* required: randomly assigned among 1 and 5 time units;
3. *Kind of deadline*: randomly assigned between hard (cannot be postponed) and soft deadline (can be postponed);
4. *Priority*: randomly assigned among 1 and 5 (the greater the value, the greater the priority).

Table 1 reports the number of agents per category. The choice of agent distribution was guided by a preliminary context analysis. Generally, in a hospital, the number of patients is greater than the number of nurses, which is greater than the number of doctors. As the National Nurse United association reports (https://www.nationalnursesunited.org/ratios, accessed on 13 December 2021), the recommended nurse-to-patient ratio should range from 1:1 to 1:4. Similarly, the doctor-to-nurse ratio fluctuates from 1:2 to 1:4 (https://www.oecd.org/coronavirus/en/data-insights/number-of-medical-doctors-and-nurses, accessed on 13 December 2021). Certainly, these numbers undergo various fluctuations, both by country and by department. Therefore, in order to stick to the proportions identified, we choose to consider one doctor, three nurses, and five patients. As far as it concerns devices, the most important factor is not how many devices there are but what the system load is. Thus, for

the sake of simplicity, the system is sized thinking of a human–device ratio equal to 1. This allows us to better interpret the effect of the request probability.

Table 1. Number of agents for each category.

	E_1	E_2	E_3	D	N	P
Number	3	3	3	1	3	5

Table 2 reports the average performance of the categories for each task. These values, together with a standard deviation of 20%, are used to generate the average performance of each category member; i.e., this dimension determines the competence of the agents in the various tasks. Of course, this is an objective and internal property of the agent. As such, this value cannot be accessed directly but only subjectively estimated. Unlike competence, willingness values are determined only by the specific system conditions. Of course, in general, an agent has an inherent predisposition to be willing/unwilling to perform a task for a specific trustor. However, in this specific case, we assume that agents are always well disposed toward other partners. Therefore, the only limitations in terms of willingness are due to the effective availability of the agents, i.e., whether or not they are engaged in other tasks.

Table 2. Average performance of the category members for each task.

	E_1	E_2	E_3	D	N	P
τ_1	100%	50%	75%	75%	75%	50%
τ_2	75%	100%	75%	75%	75%	50%
τ_3	50%	75%	75%	75%	75%	50%
τ_{doctor}	0%	0%	0%	90%	0%	0%
τ_{nurse}	0%	0%	0%	0%	90%	0%
$\tau_{patient}$	0%	0%	0%	0%	0%	90%

At each time unit, with a probability *requestProbability*, each agent generates a task request. Then, it ranks all the possible partners in the system based on their availability, physical proximity (except for τ_2), and competence. It inquiries the potential partners as long as a trustee accepts its request. As far as it concerns trustee's behavior, since our goal is to show the potential of artificial devices, we considered simpler strategies for humans and more elaborate decisions for devices. So, for the sake of simplicity, a busy human trustee just rejects further requests. On the other side, an artificial device decides how to act according to the strategies described in Section 2.

As far as it concerns trustworthiness assessment, despite the very rich literature about this topic, there is no standard solution to solve this problem. First of all, there are many theoretical models in the literature. We refer to the model proposed by Castelfranchi and Falcone [28], which identifies the core trust in the two components of competence and willingness (see Equation (1)).

$$Trustworthiness = f(Competence, Willingness) \quad (1)$$

Nevertheless, these models must be instantiated in the specific domains. As far as we are concerned, we represent trust operationally with a value in [0, 1], where 0 means absolute distrust and 1 means maximum trust. It depends on various components. In other words, we need to discuss the nature of the function f. The model proposed by Castelfranchi–Falcone is general, so it can be instantiated in different formulations, be they linear or non-linear. We choose to refer to linear models. Due to their intrinsic characteristics and potential, linear models have been widely used within social science to reproduce mental processes. More in detail, they allow us to study a particular behavior or social phenomenon, relating it to different cognitive variables or environmental factors

that influence or determine it. For example, it is possible to model the propensity to trust. Among the many works, in [35], the authors investigate the links between interpersonal trust and competences, relations, and cooperation in Polish telecommunications companies. As a further example, in [36], a variation of linear regression analysis is considered to model trust ratings of delivered services as a function of QoS (Quality of Services). Therefore, it remains to be discussed what weight to assign to the two variables, competence and willingness. They do not necessarily have the same importance. For example, there may be contexts in which willingness is taken for granted. Furthermore, since trust is a subjective dimension, it certainly depends on the trustor, who may consider one aspect more important than the another. The same stands for the trustee, the context, etc. However, in our specific case, both have an effect on the trustworthiness of the agents, and we have no reason to prefer one component over the other. For this reason, we choose to give them equal importance.

The two sub-dimensions of competence and willingness are computed each time a task is completed, interrupted, or refused. Equation (2) shows how competence is updated. Specifically, *OldCompetence* represents the belief of the trustor about the competence of the trustee before requiring the execution of the new task. At the beginning of the simulation, rather than starting from a situation of complete uncertainty, its initial value is determined from the trustee's category of belonging. $Performance \in [0, 1]$ represents the task outcome, i.e., how good was the trustee in executing the assigned task. Resuming, *Competence* is updated as the weighted mean of its precedent value *OldCompetence* and the recent *Performance*. In a similar way, in Equation (3), *OldWillingness* represents the estimation of willingness before the execution of the new task. Its initial value is equal to 1, since we agents are supposed to be always willing to accept task requests. Concerning *Availability*, it is a Boolean variable, and it is equal to 1 if the trustee successfully accomplishes the task or 0 if it refuses/interrupts the task. In Equations (2) and (3), $\alpha = 0.7$ and $\beta = 0.3$ were set to give more importance to past experience and to avoid excessive fluctuations in trustworthiness values. In Equation (4), *Trustworthiness* is computed as the average value of *Competence* and *Willingness*. It is important to underline that the values of competence, willingness, and trustworthiness are always linked to a trustor, a trustee, a task, and a context. For simplicity of reading, we omit such indices.

$$Competence = OldCompetence \times \alpha + Performance \times \beta \tag{2}$$

$$Willingness = OldWillingness \times \alpha + Availability \times \beta \tag{3}$$

$$Trustworthiness = \frac{Competence + Willingness}{2} \tag{4}$$

After analyzing the evolution of the model in a reference scenario, we compared the results of three possible cases:

1. *Trustworthiness-based* selection: the trustor selects its partners according to their trustworthiness. This case is used as a reference scenario;
2. *Random-based* selection: the trustor does not know other agents' trustworthiness; thus, it chooses its partners randomly;
3. *Trust-based* selection: the trustor selects its partners according to their trustworthiness. Additionally, trustees are able to evaluate trust; thus, they can introduce a further optimization to the system.

Our analysis starts considering the trustworthiness scenario, since it allows us to illustrate the whole functioning of the platform.

To evaluate the simulation results and investigate the effects of the proposed interaction modes, let us consider the following dimensions:

1. Percentage of *completed*, *interrupted*, and *refused* tasks;
2. Average trustworthiness value *tw* for the relationships H2M, M2M, and M2H;
3. Average device *performance* in the cases H2M and M2M.

In this first scenario, let us verify how the proposed interaction model works when varying:

1. *Request probability*: set to 10% or 25%;
2. *Error probability*: from 0% to 25%.

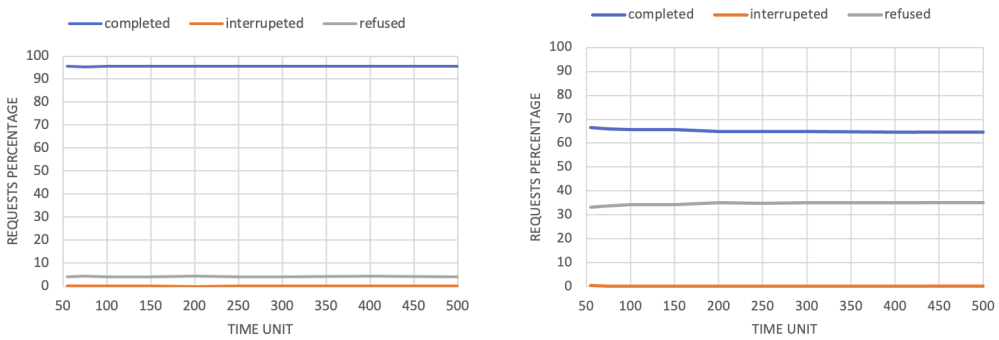
The values of 10% and 25% have been chosen in order to represent, respectively, a situation of unload and overload for the network. On average, agents take 3 time units to perform a task. With a 10% request probability, a device receives on average one task every 5 time units (half from humans and half from other devices), so the estimated average load is 60%. On the other hand, with a 25% request probability, a device receives a task every 2 time units; therefore, the estimated average load is 150%. Therefore, we need the first scenario to investigate the behavior of the model in an ideal case, that is, when the devices are almost always available. The second scenario, on the other hand, represents a more interesting situation, in which the devices have to manage conflicting situations. Of course, these are average expected values, generated randomly, so it is possible that moments of greater loading or unloading may occur.

A preliminary analysis of the framework highlighted the presence of strong fluctuations in the output values at the beginning of the simulation. This is due to the fact that when the simulation starts, the agents do not possess precise knowledge on the actual capabilities and availability of their partners as well as the system load. To overcome such problems and to eliminate the random effects on the output, we considered a transient phase equal to 50 time units. Indeed, this time window allows us to analyze stable results without such variability.

4. Results

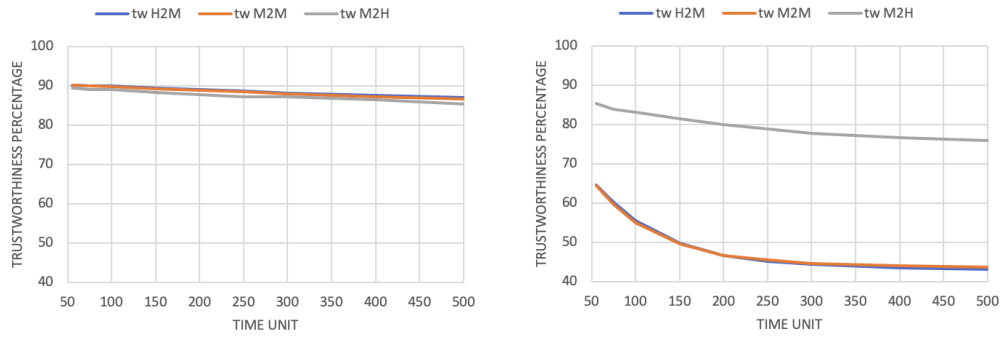
Regarding request probability, it strongly influences the performance of the system. In fact, as requests increase, it is impossible for the agents to satisfy all trustors (Figure 2a,b). In the 10% case, agents manage to handle basically all requests, while in the latter case, the percentage of refused tasks increases at the expense of completed tasks.

As a consequence, trustors will start perceiving trustees as less trustworthy because of their lower availability. Figure 3a,b show the average value of trustworthiness, respectively, with a 10% and 25% request probability. Note that the M2H value decreases less because users receive half of the requests compared to devices (H2M and M2M). Moreover, the evaluations of trustworthiness are the same in the H2M and M2M cases. Indeed, both dimensions evaluate the same thing (the trustworthiness of the devices as task executors) but from different points of view (human or device trustor).



(**a**) 10% request probability (**b**) 25% request probability

Figure 2. Percentage of completed, interrupted, and refused requests in the trustworthiness-based scenario.



(**a**) 10% request probability (**b**) 25% request probability

Figure 3. Average trustworthiness evaluation in the trustworthiness-based scenario.

4.1. The Effect of Error

In this case, we verify how errors influence the performance of the system. As mentioned in the previous sections, when a device identifies an error in the task execution, the device generates a request for help, i.e., it requests the execution of a new subtask to a human agent. The result of this subtask, necessary to solve the error, determines the possibility of completing the main task. In case no human is available for help, the original task fails.

If there had not been the possibility of managing the tasks affected by the error, they would fail. Consequently, the devices would be perceived as less trustworthy. On the contrary, the results in Figure 4 show that as the error increases, the trustworthiness of the devices increases, too. Of course, while handling errors has a positive effect on the system, the presence of error itself introduces drawbacks (Figure 5). As regards human agents, they will have to carry out a greater number of tasks, which results in an M2H trustworthiness decrease: since humans are busy managing tasks instead of devices, they are less willing to accept new ones. This means that (1) given that human agents are less available, the number of unaccepted requests increases, and (2) the number of interrupted requests increases, since in case of error, human agents are not able to help devices.

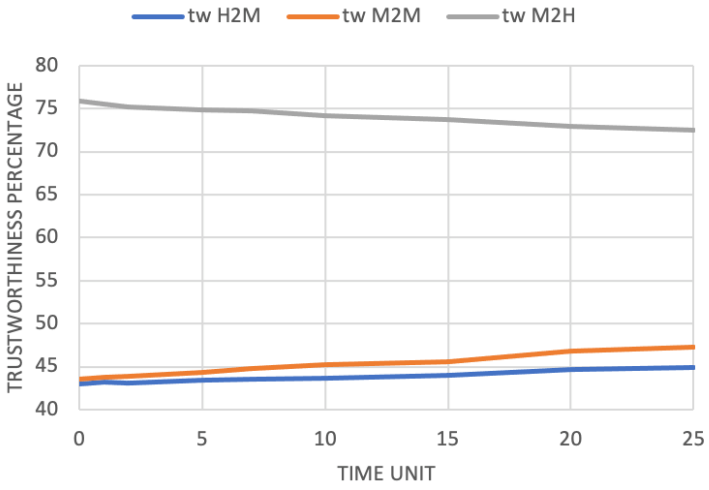


Figure 4. Average trustworthiness evaluation in the trustworthiness-based scenario after 500 time units with 25% request probability, as the error probability changes.

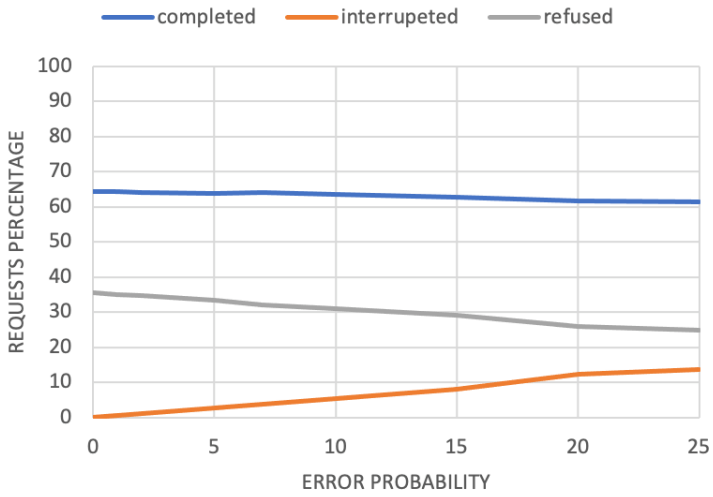
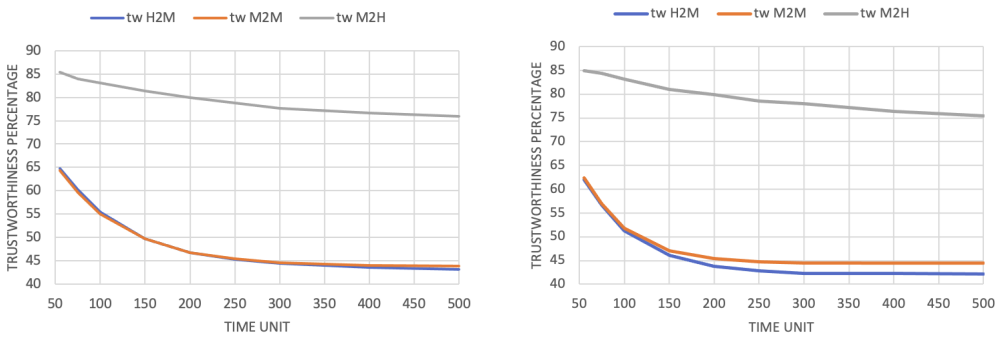


Figure 5. Percentage of completed, interrupted, and refused requests in the trustworthiness-based scenario after 500 time units with 25% request probability as the error probability changes.

4.2. Random-Based and Trust-Based Scenarios

In this further experiment, we are interested in identifying and quantifying the effect that trustworthiness (the trustor's evaluation of the trustee) and trust (the trustee's estimation of the trustor's belief about its trustworthiness) have on the system. Referring to the setting considered above (25% request probability), we compared these two approaches with the random-based one.

As in the other cases, let us start our analysis considering the average trustworthiness evaluation. Figure 6a,b reports the trustworthiness values for the trustworthiness-based and trust-based cases. Of course, the random-based scenario is not considered in this picture, as agents do not evaluate their partners in that case. Paradoxically, the two figures show us that in the trust-based scenario, agents are perceived as less trustworthy compared to the trustworthiness-based case. Such an outcome is different from what one would expect to find, since although agents behave in a more reliable way, they are trusted less. What may seem like a performance worsening is, even if counter intuitively, a positive phenomenon. Indeed, introducing trust prevents the trustworthiness of the trustees from being overestimated by the trustors. While this causes agents to be perceived as less trustworthy in the trust-based scenario, it also leads to a better partner selection.

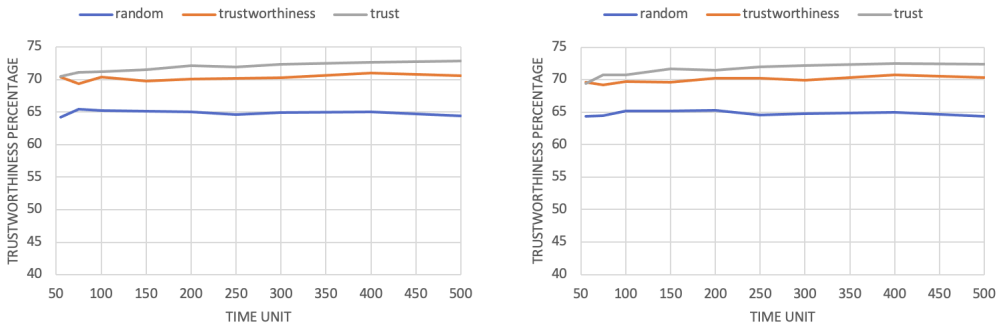


(**a**) Trustworthiness-based scenario (**b**) Trust-based scenario

Figure 6. Average trustworthiness evaluation with 25% request probability.

Thus, to verify the presence of any improvement, it is necessary to take into account the outcome of the interactions in the different scenarios. Figure 7 shows the average performance received by trustors in the H2M (Figure 7a) and M2M (Figure 7b) relationships. Indeed, results confirm an increase of the average performance ranging from 4 to 6% when

we introduce trustworthiness and from 6 to 8% when even trust is considered. Given that we are dealing with average values, this is a remarkable result.



(**a**) H2M relationship (**b**) M2M relationship

Figure 7. Comparison of average performance among random-based, trustworthiness-based, and trust-based scenarios with 25% request probability.

5. Discussion

The results of the simulation highlight some of the advantages that follow from the proposed modeling. The first one is that when agents are busy, instead of refusing new task requests, they can consider alternative strategies. Agents can reorganize task execution, postponing or reassigning them. The effectiveness of such possibility has already been recognized in M2M models but not in H2M ones. In particular, reorganizing tasks represents a further added value because even the trustees are themselves trustors. As a consequence of their knowledge of the others' ability and their possibility to handle the assignment of two tasks at the same time, agents are able to further optimize the final performance received by trustors.

Then, we introduced a simulation study to test and show the functioning of the proposed system. Indeed, the results obtained in the simulations provide several interesting food for thought.

We started analyzing the effect of the request probability and error probability. As for the first, increasing the request probability has the effect of reducing the devices' trustworthiness. In fact, although correctly identified as competent, trustees are less available toward their interlocutors. In other words, the dimension of willingness is negatively affected by an excessive load of requests (a more analytic evaluation should separate the direct and indirect responsibility with respect to this devices' willingness).

An interesting point is that of error management. It is worth underlining that by error, we mean a situation external to the trustee, not linked to its intrinsic ability, which prevents it from completing the task. In general, in H2M models, when an error arises, the devices have no choice but to let the tasks fail. This is due to the limits of their model of interaction, which does not provide them with the proper instruments to solve such problems. This situation can also occur in M2M models. In fact, since errors go beyond the actual capabilities of the devices, involving an additional device may not represent an effective solution. Even in this case, the framework introduced allows the device to find a valid alternative to face the situation. In these cases, the devices can request help from human agents to overcome the error, since humans have different characteristics and capabilities compared to devices. When the error has been tackled, devices can continue with the execution of their original task. Indeed, analyzing the effect of error management, it appears that with the same number of requests, H2M and M2M trustworthiness increase. There are two reasons for this positive outcome. On the one hand, since they prevent errors from causing tasks to fail, there is no negative feedback on the performance of the devices. Moreover, given that a percentage of their tasks is reassigned to humans, devices will be more available to accept new tasks.

Lastly, we compared the three different approaches of partner selection in order to verify how they affect the different relations: random-based, trustworthiness-based, and trust-based. Of course, the random-based approach has just been considered as a reference case, and its comparison with the trustworthiness-based approach has only been used as a term of comparison. Indeed, the effectiveness of the trustworthiness analysis is a consolidated result in the current literature. As far as it concerns trust, its introduction has basically two effects. First of all, trustees are perceived as less trustworthy than in the trustworthiness-based case. This happens because since trustees know their trustworthiness better than the other agents, they have the possibility to correct the trustor's belief in case this one overestimates their actual possibilities.

By contrast, while the evaluation of trustworthiness decreases, the average performance received by (human and artificial) trustors increases. Remarkably, the comparison between the trustworthiness-based and the trust-based scenarios allows us to highlight a further point: it is true that systems should aim to be perceived as reliable as possible, but this must be done by behaving in the most reliable way possible. The main purpose of artificial systems should not be to maximize the user's perception of them. While even this is an important topic, they should above all aim to improve the actual performance provided to the user. This also means being willing to be perceived as less trustworthy in order to maximize user utility.

6. Conclusions

In this work, we start from the consideration that, in order to better evaluate the potential of HMI models and to overcome the current limitations, it is fundamental to look simultaneously at the different types of interaction that arise between human and artificial agents. Indeed, many works have contributed to the creation of trust models in the field of human–robot or human–AI interaction. Regrettably, most of the existing approaches focus just on a specific dimension of the trust relationship (trust or trustworthiness). This contribution tried to tackle this research gap by proposing an integrated approach of trust and trustworthiness distributed on the various (human or artificial) actors of the interaction. Indeed, our approach confirms the initial claim, since taking into account different kinds of interaction at the same time allows to take advantage of the individual benefits they entail while also introducing new possibilities.

It is important to underline, as a limitation, that this work took into consideration a linear implementation of trust, which assigned equal weight to the components of competence and willingness. As future work, we are also interested in evaluating the impact of these weights on agents' decisions.

To conclude, the results of this work, albeit more theoretical, provide interesting insights for the evolution of HMI models. As future research, we are currently planning a field study involving human users in their actual interaction with robots. Such study will allow us to provide quantitative estimates of the effects of the introduced model on the dynamics of the interaction.

Author Contributions: Conceptualization, A.S., F.C. and R.F.; Formal analysis, A.S., F.C. and R.F.; Methodology, A.S. and R.F.; Project administration, R.F.; Software, A.S.; Supervision, R.F.; Writing and original draft, A.S., F.C. and R.F.; Writing, review and editing, A.S., F.C. and R.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

HMI	human–machine interaction
HRI	human–robot interaction
MAS	multi-agent systems
IoT	Internet of Things
ToM	Theory of Mind
H2M	trust relationship involving a human trustor and an artificial entity as trustee
M2M	trust relationship involving artificial entities as trustor and trustee
M2H	trust relationship involving an artificial entity as trustor and a human as trustee

References

1. Hou, M.; Ho, G.; Dunwoody, D. IMPACTS: A trust model for human-autonomy teaming. *Hum.-Intell. Syst. Integr.* **2021**, *3*, 79–97. [CrossRef]
2. Falcone, R.; Sapienza, A. On the users' acceptance of IoT systems: A theoretical approach. *Information* **2018**, *9*, 53. [CrossRef]
3. Di Dio, C.; Manzi, F.; Peretti, G.; Cangelosi, A.; Harris, P.L.; Massaro, D.; Marchetti, A. Shall I trust you? From child–robot interaction to trusting relationships. *Front. Psychol.* **2020**, *11*, 469. [CrossRef] [PubMed]
4. Pinyol, I.; Sabater-Mir, J. Computational trust and reputation models for open multi-agent systems: A review. *Artif. Intell. Rev.* **2013**, *40*, 1–25. [CrossRef]
5. Ekman, F.; Johansson, M.; Sochor, J. Creating appropriate trust in automated vehicle systems: A framework for HMI design. *IEEE Trans. Hum.-Mach. Syst.* **2017**, *48*, 95–101. [CrossRef]
6. Morra, L.; Lamberti, F.; Pratticó, F.G.; La Rosa, S.; Montuschi, P. Building trust in autonomous vehicles: Role of virtual reality driving simulators in HMI design. *IEEE Trans. Veh. Technol.* **2019**, *68*, 9438–9450. [CrossRef]
7. Chen, R.; Guo, J.; Bao, F. Trust management for SOA-based IoT and its application to service composition. *IEEE Trans. Serv. Comput.* **2014**, *9*, 482–495. [CrossRef]
8. Guo, G.; Zhang, J.; Yorke-Smith, N. A novel recommendation model regularized with user trust and item ratings. *IEEE Trans. Knowl. Data Eng.* **2016**, *28*, 1607–1620. [CrossRef]
9. Zhong, Y.; Bhargava, B.; Lu, Y.; Angin, P. A computational dynamic trust model for user authorization. *IEEE Trans. Depend. Secur. Comput.* **2014**, *12*, 1–15. [CrossRef]
10. Jayasinghe, U.; Lee, G.M.; Um, T.W.; Shi, Q. Machine learning based trust computational model for IoT services. *IEEE Trans. Sustain. Comput.* **2018**, *4*, 39–52. [CrossRef]
11. Messina, F.; Pappalardo, G.; Rosaci, D.; Santoro, C.; Sarné, G.M. A trust-aware, self-organizing system for large-scale federations of utility computing infrastructures. *Future Gener. Comput. Syst.* **2016**, *56*, 77–94. [CrossRef]
12. Ba-hutair, M.N.; Bouguettaya, A.; Neiat, A.G. Multi-perspective trust management framework for crowdsourced IoT services. *IEEE Trans. Serv. Comput.* **2021**. [CrossRef]
13. Fortino, G.; Messina, F.; Rosaci, D.; Sarné, G.M. Using trust and local reputation for group formation in the cloud of things. *Future Gener. Comput. Syst.* **2018**, *89*, 804–815. [CrossRef]
14. Wang, T.; Luo, H.; Jia, W.; Liu, A.; Xie, M. MTES: An intelligent trust evaluation scheme in sensor-cloud-enabled industrial Internet of Things. *IEEE Trans. Ind. Inform.* **2019**, *16*, 2054–2062. [CrossRef]
15. Hu, W.L.; Akash, K.; Reid, T.; Jain, N. Computational modeling of the dynamics of human trust during human–machine interactions. *IEEE Trans. Hum.-Mach. Syst.* **2018**, *49*, 485–497. [CrossRef]
16. Xu, A.; Dudek, G. Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In Proceedings of the 2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Portland, OR, USA, 2–5 March 2015; pp. 221–228.
17. Edmonds, M.; Gao, F.; Liu, H.; Xie, X.; Qi, S.; Rothrock, B.; Zhu, Y.; Wu, Y.N.; Lu, H.; Zhu, S.C. A tale of two explanations: Enhancing human trust by explaining robot behavior. *Sci. Robot.* **2019**, *4*, eaay4663. [CrossRef]
18. Nikolaidis, S.; Kuznetsov, A.; Hsu, D.; Srinivasa, S. Formalizing human-robot mutual adaptation: A bounded memory model. In Proceedings of the 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), Christchurch, New Zealand, 7–10 March 2016; pp. 75–82.
19. Chen, M.; Nikolaidis, S.; Soh, H.; Hsu, D.; Srinivasa, S. Trust-aware decision making for human-robot collaboration: Model learning and planning. *ACM Trans. Hum.-Robot Interact.* **2020**, *9*, 1–23. [CrossRef]
20. Cantucci, F.; Falcone, R. Towards trustworthiness and transparency in social human-robot interaction. In Proceedings of the 2020 IEEE International Conference on Human-Machine Systems (ICHMS), Rome, Italy, 7–9 September 2020; pp. 1–6.
21. Nikolaidis, S.; Hsu, D.; Srinivasa, S. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *Int. J. Robot. Res.* **2017**, *36*, 618–634. [CrossRef]
22. Hannum, C.; Li, R.; Wang, W. Trust or Not?: A Computational Robot-Trusting-Human Model for Human-Robot Collaborative Tasks. In Proceedings of the 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, 10–13 December 2020; pp. 5689–5691.

23. Vinanzi, S.; Patacchiola, M.; Chella, A.; Cangelosi, A. Would a robot trust you? Developmental robotics model of trust and theory of mind. *Philos. Trans. R. Soc. B* **2019**, *374*, 20180032. [CrossRef]
24. Hu, Y.; Abe, N.; Benallegue, M.; Yamanobe, N.; Venture, G.; Yoshida, E. Toward Active Physical Human–Robot Interaction: Quantifying the Human State During Interactions. *IEEE Trans. Hum.-Mach. Syst.* **2022**. [CrossRef]
25. Falcone, R.; Sapienza, A.; Castelfranchi, C. The relevance of categories for trusting information sources. *ACM Trans. Internet Technol.* **2015**, *15*, 13. [CrossRef]
26. Liu, X.; Datta, A.; Rzadca, K. Trust beyond reputation: A computational trust model based on stereotypes. *Electron. Commer. Res. Appl.* **2013**, *12*, 24–39. [CrossRef]
27. Mehrotra, S. Modelling Trust in Human-AI Interaction. In Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems, Virtual, 3–7 May 2021; pp. 1826–1828.
28. Castelfranchi, C.; Falcone, R. *Trust Theory: A Socio-Cognitive and Computational Model*; John Wiley & Sons: Hoboken, NJ, USA, 2010; Volume 18.
29. Falcone, R.; Castelfranchi, C. Social trust: A cognitive approach. In *Trust and Deception in Virtual Societies*; Springer: Dordrecht, The Netherlands, 2001; pp. 55–90.
30. Castelfranchi, C.; Falcone, R. Principles of trust for MAS: Cognitive anatomy, social importance, and quantification. In Proceedings of the International Conference on Multi Agent Systems (Cat. No. 98EX160), Paris, France, 3–7 July 1998; pp. 72–79.
31. Falcone, R.; Sapienza, A.; Castelfranchi, C. Recommendation of categories in an agents world: The role of (not) local communicative environments. In Proceedings of the 2015 13th Annual Conference on Privacy, Security and Trust (PST), Izmir, Turkey, 21–23 July 2015; pp. 7–13.
32. Conte, R.; Paolucci, M. *Reputation in Artificial Societies: Social Beliefs for Social Order*; Kluwer Academic Publishers: Boston, MA, USA, 2002.
33. Sapienza, A.; Falcone, R. Evaluating agents' trustworthiness within virtual societies in case of no direct experience. *Cogn. Syst. Res.* **2020**, *64*, 164–173. [CrossRef]
34. Burnett, C.; Norman, T.J.; Sycara, K. Stereotypical trust and bias in dynamic multiagent systems. *ACM Trans. Intell. Syst. Technol.* **2013**, *4*, 1–22. [CrossRef]
35. Bulińska-Stangrecka, H.; Bagieńska, A. Investigating the links of interpersonal trust in telecommunications companies. *Sustainability* **2018**, *10*, 2555. [CrossRef]
36. Mashinchi, M.H.; Li, L.; Orgun, M.A.; Wang, Y. The prediction of trust rating based on the quality of services using fuzzy linear regression. In Proceedings of the 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), Taipei, Taiwan, 27–30 June 2011; pp. 1953–1959.