Trustful Action Suggestion in Human Agent Interaction

[Extended Abstract]

Nuno Xu
Instituto Superior Técnico
Lisbon, Portugal
nuno.xu@tecnico.ulisboa.pt

ABSTRACT

Trust is an essential ingredient for cooperation and collaboration, so if we want to further develop autonomous collaborative agents, we must address the issue of trust in such relationships. For that reason, computational trust in Human-Robot Interaction (HRI) has seen a great spike of interest in recent years, however the literature has been only focused in issues like design, animation and modelling. This thesis addresses the uncharted matter of actively improving trust by suggesting trustful actions to the agent. Towards this goal we developed a Cognitive Trust Model capable of representing a belief based Trust model and suggest utterances with the goal of improving Trust. Furthermore we also designed a novel Trust and Rapport evaluating scenario, Quick Numbers, to be able to properly evaluate the combined effect of ability and willingness in Trust and have a simple measurement of Trust embedded in the testing scenario. Finally, we describe the User Studies to evaluate the Trust Model using the Quick Numbers scenario.

Keywords

Artificial Intelligence (AI), HRI, Trust Modelling, Trust Evaluation

1. INTRODUCTION

Trust has been described in Psychology as being one of the most important components of interpersonal relationships [34]. It is undeniable the need of trust to promote cooperation and collaboration between two parties, specially regarding who should one trust and what is worth entrusting. So as AI research gravitates towards the development of Intelligent Agent Systems [27], a focal concern is the performance of collaborative tasks [13, 2]. And while the amount of literature has been increasing, we found it surprising that not enough work has been done in Human-Agent Interaction (HAI) focusing on trust, other than on design issues [6] and the sub-field of HRI [11, 35], specially when so much has been done regarding Trust in Automation (TiA) [18, 19]. This reveals that while the area has so much potential, the level of understanding is still very shallow, only deeply focused in certain areas [12].

Multi-Agent System (MAS) Trust and Reputation modelling is one of the areas that has been having a great increase of interest lately, specially ever since the advent of Peer-To-Peer (P2P) e-commerce in platforms like *eBay*[1]. For this applications, tools and solutions to ensure trust were needed for a new reality of a mass amount of anonymous entities constantly entering and exiting the environment and performing trading transactions through an open space. However almost all research focuses purely on the creation and maintenance of a trust model about the environment around the agent, providing a rank for other agents, but not taking into account the agent's own stance in the environment. Additionally most of this models' designs are based in statistical and game theoretical concepts [12] which makes them difficult to understand, analyse and, most importantly, describe their evaluative reasoning in a human understandable manner. Castelfranchi and Falcone [8] tried to solve these problems with the introduction of cognitive models, by mapping the trust model to the agent's mental state, composed by beliefs and goals, very akin to existing cognitive agent architectures like Belief-Desire-Intention (BDI) [25]. Then some systems, like Repage [28], created implementations of this new paradigm of trust modelling, where most of the models were purely theoretical.

Nevertheless, there is a gap in this area of research that we wish to address with our work: the lack of an implementation for an action suggester based on the agent's trust model, with the goal to improve the strength of our beliefs in the model and to improve trust in our agent. While one could argue that this is the responsibility of the decision making or planner component of the agent, we believe that a dedicated module will ease the complexity of decision by making it more modular, and also allowing for the trust model to take a more active part in the decision making process. To our knowledge, no attempts have been done towards this goal, so we propose to develop a Trust Model that: firstly, is capable of creating a cognitive model representing the mental state of the user's trust in the agent, following Castlefranchi and Falcone's concepts of Cognitive Trust Modelling and taking inspiration from Repage's architecture, and secondly, able to suggest what actions should be used to improve trust on the agent.

Developing this model also provided the opportunity to address Trust evaluation, as we found a lack of scenarios in HAI that would address Trust's two main components, Ability and Willingness, simultaneously. This urged us to design a scenario that would address this issues and remain relevant to other studies in this area. The scenario was developed in collaboration with Henriques' thesis work on *Rapport - Establishing Harmonious Relationship Between Robots and Humans* [15].

[1]eBay Auctions: http://www.ebay.com/

2. BACKGROUND

Before discussing related work and our solution to the problem, we will present the main concepts that will be mentioned in the rest of this paper, specifically regarding trust and reputation.

2.1 Trust

Trust is regarded throughout the literature as one of the fundamental components of human society, being essential in cooperative and collaborative behaviour, having been studied in a multitude of disciplines, from Psychology and Sociology, to Philosophy and Economy [26, 18, 30]. For that reason, it is no wonder that it acquired a very large number of different definitions throughout the years of study, causing the problem of not existing a consensus on a definition of trust [9]. In the scope of our work, the most relevant start for our discussion is the dyadic definition of trust: 'an orientation of an actor (the **truster**) toward a specific person (the **trustee**) with whom the actor is in some way interdependent' (taken from [34]), as we want to focus on interpersonal relationships. This definition has been expanded throughout the literature, often adapted to fit the context or scope of the work, but three main definitions are highlighted in computational trust:

- First, Gambetta [10] defined trust as follows: "Trust is the *subjective probability* by which an individual, A, *expects* that another individual, B, performs a given action on which its *welfare depends*" (taken from [9]). This is accepted by most authors as one of the most classical definitions of trust, but it is too restrictive with its uni-dimensionality, as it only refers to predictability of the trustor, and does not take into account competence in executing the given action.

- Marsh [21] was the first author to formalize trust as a measurable Computational Concept, continuing the perspective of reducing trust to a numerical value, set by Gambetta [10], but also adding that: X trusts Y if, and only if, "X *expects* that Y will behave according to X's best interest, and will not attempt to harm X" (taken from [9]). This definition does not represent other parts of trust, such as the notion that trustor must ascertain some risk from delegating the action to the trustee.

- Castelfranchi and Falcone then introduced a Cognitive aspect to Computational Trust [8]. They define trust as the mental state of the trustor and the action in which the trustor refers upon the trustee to perform. This is the definition of trust that we will adopt throughout the rest of the report, as it represents a vision of trust that takes into account the trustor set of beliefs and intentions, approaching it to an agent's cognitive model, while also linking trust to the action being performed, as one might trust another for certain types of actions and not for others (e.g. I may trust my squire to polish my sword, but not to swing it).

2.1.1 Castelfranchi and Falcone's Trust

More explicitly, Castelfranchi and Falcone [8] state that trust is a conjunction of three concepts:

- A *mental attitude* or (pre)disposition of the agent towards another agent; this is represented by beliefs about the trustees' qualities and defects;

- A *decision* to rely upon another, and therefore making the trustor "vulnerable" to the possible negative actions of the trustee;

- The *act* of trusting another agent and the following behaviour of counting on the trustee to perform according to plan.

By describing trust as a mental attitude it is also implied that: "Only a cognitive agent can trust another agent; only an agent endowed with goals and beliefs" [9].

From this definition we should also address one important component, **Delegation**, which happens when an agent (X) needs or likes the action delegated to another agent (Y), so X includes it in his plans, therefore relying on Y. X plans to achieve his goal through Y. So, he formulates in his mind a multi-agent plan with a state or action goal being Y's delegated [8].

2.2 Reputation and Image

Reputation is also a concept that appears very often linked with trust in the literature, specially since recent models created for representing trust have been focused on MASs (see [1, 29, 28, 17, 23]), where most have also been developed to include reputation as a source of trust.

An agent is not only influenced by their own beliefs about the subject, the *Image*, but also by what other agents say about it, its *Reputation*.

We describe Image and Reputation by Sabater's definition in [28]: Image is defined as the agent's personal belief about a certain property of the target agent, be it a physical, mental or social trait. Reputation is a meta-belief about an impersonal evaluation of the target, in other words, it is the belief on the evaluation being circulated about the target. On a more concrete level, reputation is separated between *shared evaluation* and *shared voice*. Consider that an agent has beliefs about how other agents evaluate a certain target, if in a set of agents these beliefs converge to a value (e.g. "good" or "bad") we can say that there exists a shared evaluation of the target. It is important to note that all voice sharing agents are known and well defined. A shared voice is a belief that another set of agents themselves believe that an evaluation of the target exists. In other words, it is the belief that a group of agents will consistently report that a voice exists. These meta-beliefs are considered important as one is not required to believe that other's evaluation is correct, but might still believe that it exists.

The mental decisions regarding reputation can be categorized as follows:

- Epistemic decisions: accepting trust beliefs to update or generate a given image or reputation;

- Pragmatic-Strategic decisions: using trust beliefs to decide how to behave towards other agents;

- Memetic decisions: transmitting trust beliefs to others.

This difference of possible decisions allows to describe how one may transmit reputation without having the responsibility for the credibility or truthfulness of the content transmitted, as one does not have to commit to accepting the reputation value, and just say that the rumour exists.

2.3 Game Theory

Game Theory is the field of study that defines and analyses situations involving conflict or cooperation between multiple intelligent decision makers. These situations are called a game, and they are distilled to their core argument, by defining the limited and simple set of actions that the players may perform, and how do they affect the players. It then analyses the decision strategies for each player, by assuming that both will try to maximise their payoff (how much the player gains) with their action.

3. RELATED WORK

Computational Trust research has been focused on modelling trust in MASs, specially on open e-commerce environments [12, 14, 24, 22, 16], with at least 106 models created [12], since the formalization of trust as a measurable property by Marsh in 1994 [21]. We will present some trust models from which we will take inspiration while creating our own, and some work done in measuring trust in HRI.

3.1 Trust Models

For related work concerning Trust Models we will focus on **Cognitive** Trust Models, first introduced by Castelfranchi and Falcone [8], which are defined by measuring trust on the strength of an agent's beliefs and the changes enacted through the consequent act of trusting. We focused on modelling trust through multiple dimensions, with the intent of having trust depend on the action to perform, context and agent performing the task and having these dimensions represented explicitly in the model, something that it is not possible with **Numerical** models, like the one introduced by [21].

3.1.1 Castelfranchi and Falcone's model

Having developed the concept of Cognitive Trust Models, this author's model is generally regarded as a classical basis for most other authors, and while we do not use the entirety of the concepts defined in this model, it is worth describing, as it was also a source of inspiration to other authors referenced in this paper. The model is characterised around their definition referred in Section 2.1.1, through a central core, composed by a five-part relation, between:

- The trustor (**X**);
- The trustee (**Y**);
- The context where they are inserted in (**C**);
- A task ($\boldsymbol{\tau}$) defined by the pair (α, ρ), where $\boldsymbol{\alpha}$ is the action entrusted to the trustee, that possibly produces an outcome $\boldsymbol{\rho}$, contained in the goal of X (g_x);
- The goal of the trustor ($\boldsymbol{g_x}$).

More shortly represented by equation 1.

$$TRUST(X\ Y\ C\ \tau\ g_x) \tag{1}$$

This defines Trust as goal-oriented, contextual, and multi-dimensional, as from the point of view of the trustor, it varies not only on the trustee, but also from the overall context, the action that is being delegated, and the particular goal of the trustor. For example, if the goal of the trustor is simple to perform and not very critical to him, he may be more willing to delegate the task, and trust another agent to perform such task. Adjustments can be attached to this core adjusting better to the context in which it may be used. For instance, one may add an authoritative third party element to the relation in supervised security applications.

The model also conceptualizes **Expectation** as a belief of when agent X awaits for ρ to happen when an action α trusted to Y is being performed, formalized in first order logic in equation 2.

$$\begin{aligned}(Expectation\ X\ \rho) \implies (Bel_x^{t'}(\textit{will-be-true}^{t''}\rho)) \wedge \\ (Goal_x^{Period(t',t''')}(KnowWhether_X(\rho\ OR\ Not\ \rho)^{t''}))\end{aligned} \tag{2}$$

This can be used to establish what expectations the user should have in the agent, whether initial or constructed during interaction, and provide an additional measure to weight the importance of certain agent functions and actions.

As stated in the definition (Section 2.1.1) the mental attitude of the trustor X is defined by beliefs of the qualities (and faults) of Y. Therefore we can quantify the strength of our belief in a certain quality through its **Degree of Credibility (DoC)**, which is defined by a function **F** that takes all different belief sources for this quality, as shown in equation 3, where for a source sj, Str_j represents the value of the source and $Qual\text{-}i_{sjY}(\tau)$ the value of quality i of agent Y provided by the source in performing task τ.

$$\begin{aligned}DoC_X(Qual\text{-}i_{(s1,...sn),Y}(\tau)) = \\ = F_{X,Y,\tau}(Bel_X(Str_1\, Qual\text{-}i_{s1Y}(\tau)), \\ Bel_X(Str_2\, Qual\text{-}i_{s2Y}(\tau)), ..., \\ Bel_X(Str_n\, Qual\text{-}i_{snY}(\tau)))\end{aligned} \tag{3}$$

$F_{X,Y,\tau}$ associates the *strengh-of-sources (Str_j)* and *quality-values ($Qual\text{-}i_{sjY}(\tau)$)* with a probability curve. It should return a matrix with two columns, with an amount of rows corresponding to the number of quality values selected out of the received as input (since not all values must or should be used, and some may be integrated into a single value), and the first column should contain these values associated with their normalized probabilities in the second column (the probabilities sum should be 1).

For example, consider that we want agent X's DoC regarding Y's ability to clean:

- We have two sources about Y's ability to clean:
  1. X saw Y once clean quite well, but long ago, so we could attribute $Ability_{s1Y}(cleaning) = 0.8$ and $Str_1 = 0.2$;
  2. Someone X considers reliable informs that Y performed poorly recently, se we attribute $Ability_{s2Y}(cleaning) = 0.2s$ and $Str_2 = 0.6$;
- So a possible result of $DoC_X(Ability_Y(cleaning))$ is:

$$\begin{pmatrix}0.8 & 0.25\\ 0.2 & 0.75\end{pmatrix}$$

Finally **Degree of Trust (DoT)** quantifies the Trust level agent X has in Y to perform task τ according to the formula depicted in equation 4.

$$\begin{aligned}DoT_{XY\tau} = c_{Opp}\ DoC_x[Opp_y(\alpha,\rho)]\times \\ \times\ c_{Ability_y}\ DoC_x[Ability_y(\alpha)]\times \\ \times\ c_{WillDo}\ DoC_x[WillDo_y(\alpha,\rho)]\end{aligned} \tag{4}$$

Where:

- $DoC_x[Opp_y(\alpha, \rho)]$ is the DoC of X's beliefs about all contextual factors in which Y will act; in other words, the degree of Opportunity Y has to do α and result in ρ;
- $DoC_x[Ability_y(\alpha)]$ is the DoC of X's beliefs about Y's ability to perform α;
- $DoC_x[WillDo_y(\alpha, \rho)]$ is the DoC of X's beliefs concerning if Y's actually is going to perform α with the result ρ;
- c_{Opp}, $c_{Ability_y}$ and c_{WillDo} are constants representing the weight of each DoC.

This model is the most abstract, as almost all of the implementation details are left aside, particularly how the beliefs are modelled and how to or even what should be a good quantification to the quality values for the agent. This provides a lot of liberty on how to contextualize the model, and such adaptability is interesting as the model can be easily adapted to different scenarios.

3.1.2 Repage: A REPutation and ImAGE model

This system was introduced in 2006 by Sabater et al. [28] and aims to establish two different aspects to trust modelling, Image and Reputation, as defined in Section 2.2. The representation for an evaluation are fuzzy sets, defined by a tuple of five positive numbers(summing to one), where each number corresponds to a value of probability (weights) traced directly to the following scale: *very bad (VB), bad (B), neutral (N), good (G), very good (VG)*. Additionally the strength of the belief is added to the tuple, so it can be represented like this $\{w_1, w_2, ..., w_5, s\}$.

The architecture is composed by three main elements, a *memory*, a set of *detectors*, and the *analyser*. Memory is composed by predicates that are conceptually organized in different levels of abstraction and are inter-connected by a network of dependencies that propagate changes and inferences through the various predicates. The predicates contain a fuzzy evaluation belonging to one of the following types (image, reputation, shared voice, shared evaluation, valued info, evaluation from informers, and outcomes), and refer to a certain agent performing a specific role. The detectors infer new predicates, remove non-useful ones and builds the dependency network.

At the first level of the abstraction hierarchy we have the basis of information to infer predicates, *contracts*, *fulfilments* and *communication* (they are not themselves predicates, as no evaluation is attached). Contracts are agreements between two agents, while fulfilments are the results of the contract. Communication is the information about other agents that come from third parties. The second level is then constituted by inferences to an outcome, formed by a contract and its fulfilment, and valued information gathered from communications. This inferred predicates are not just tuples, they give an evaluation to the predicate, setting its belief strength.

In the next level we have two predicates: *shared voice* and *shared evaluation*. The former is inferred from communicated reputation, and the latter from communicated images.

The fourth level is composed from five types of predicates: *Candidate Image*, *Candidate Reputation*, *Image*, *Reputation* and *Confirmation*. The candidate predicates are Images and Reputations that do not have enough support yet. Special detectors turns them to fill image/reputations when a strength threshold is surpassed. Confirmation is the feedback to a communication, received from comparing it to the image of the target.

Finally the last abstractions level is composed of the predicates *cognitive dissonance* and *certainty*. Cognitive dissonance is a contradiction between relevant pieces of information that refer to the same target. This predicate may create instabilities in the mind of the individual, so the agent will most likely try to perform action in order to confirm the sources of this dissonance. Certainty represents full reliance on what the predicate asserts.

The last element is the analyser and its job is to propose actions in order to improve the accuracy of predicates in Repage and solve cognitive dissonances to produce certainty. The actions are proposed to the agent planner, leaving it to decide how to take this actions into account.

Image and Reputation are the predicates that provide a trust evaluation of a target, and as previously stated, they have a role, that represents two things: the agents interaction model, in other words, the actions that may affect to this evaluation, and a function that contextualizes the evaluative labels of *VB, B, N, G, VG*. The probability distribution of the values gives out a picture of the target interaction forecast (e.g. a probability value of 0.5 to VB gives a 50% chance of the next interaction with the target being very bad).

The work described here is the only found that tries to establish an implementable architecture for a trust model, as most of the models created are purely theoretical. Furthermore, it fits to our goals of creating a trust assessment module, corresponding to the memory and detector components, and a trust decision module, corresponding to the analyser.

4. TRUST MODEL

We sought out to develop a trust model definition that would be easily implementable, but generic enough to be able to adapt to various testing scenarios. To do this we took inspiration from the work by Sabater et al. [28] described in Section 3.1.2 by taking a similar approach to architecture where a central memory component holds the model's current state, getting updated by perceptions received from the environment. But while Repage describes a third module that suggests actions to resolve belief conflicts in the model, we instead defined such module to assume the point of view of one of the agents in the scenario and, if granted an opportune moment, it suggests actions to improve the trust relationship with a trustor. In fact, most of the design of the model was made with the intent that it would be used by one of the agents in the scenario, where the model created would be his own trust model of the world environment. And so, the model is composed by 3 main components, represented in Figure 1, and described as follows:

- **Memory**, which defines and stores the main model structure;
- **Perceptions**, a series of environment perceptions mapped to changes in the Memory;

- **Action Suggestion**, a module that outputs different actions depending on current perceptions and the state of the model.

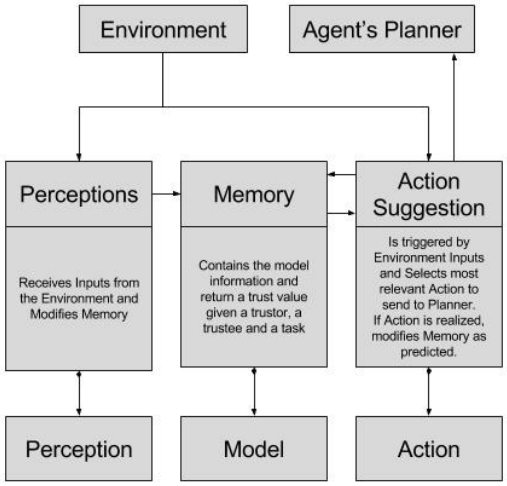

Figure 1: Model Architecture with brief descriptions, their interactions with the scenario and what they contain.

4.1 Memory

One of the main concerns while designing the model was how trust would be calculated, as we wanted to use Castelfranchi and Falcone's conceptualization of trust [9] as a basis for our definition of trust, focusing specially on it being dependent on the task entrusted, and the transferability of trust between different tasks. But starting from the five-part definition of trust, as seen in Equation 1, we decided that inserting context ($\mathbf{C}$) and the trustor's goal ($\boldsymbol{g_x}$) into the model would bring in too much complexity for the scope of this thesis, as it would require for a world state model to be kept, as well as some way to predict the trustor's goal. So we simplified, defining trust through a simpler three-part relation, involving just the trustor ($\mathbf{X}$), the trustee ($\mathbf{Y}$) and the task ($\boldsymbol{\tau}$), represented in Equation 5.

$$TRUST(X\ Y\ \tau) \tag{5}$$

So we designed the structure with the concepts and relations represented in Figure 2, and we can describe them as follows:

- **Agent**: a simple representation of a known entity in the scenario world space, serving mostly as an identifier for the entity and a container for the other agents it has information about, represented as Trustees;
- **Trustee**: each agent contains a collection of other agents it has information about, either by reputation, or by interaction, which we represent as their Trustees;
- **Trust Feature**: a piece of information a trustor has on a trustee is represented in a Trust Feature, which contains the Belief Sources of said information. The Feature Model defines and uniquely identifies what feature is represented.
- **Feature Model**: the possible set of trust features from which a trustee can be assigned is defined in a collection of Feature Models where each one uniquely identifies a possible piece of trust related information relevant to the model scenario (e.g. The trustee's ability to cook, or the willingness to drive);
- **Category**: a Feature Model must belong to a Category, making it easier to present the different type of Trust Features. This is usually intended to separate features between those relating to Ability and those related to Willingness.
- **Belief Source**: this represents a source of information on the corresponding feature, belonging to one of the 3 sub-classes depending on the origin of the information, Reputation for when reported from other agents (whether directly (e.g. talking) or indirectly (e.g. report on newspaper)), Bias for pre-existing beliefs on the feature, and Direct Contact for direct observations of the trustee, 3 values are contained to determine the associated feature's belief value:
  - Belief Value, a number between 0.0 and 1.0 describing the trustor evaluation;
  - Certainty describes how well the trustee was evaluated, in Reputation for instance, this might represent how well we trust in the reporter, and in Direct Contact how well the trustor observed the trustee performing said feature;
  - Time is just a record of when was this belief source recorded, as older records might have a lower impact in the overall belief value score, compared to newer records.
- **Task**: a representation of the possible delegation tasks in the scenario, containing the Feature Models associated with the performance of this task (e.g. The ability to serve drinks if the task is bartending). A weight is given to each Feature corresponding to its importance in the task. The various weights are normalized so that their sum is 1.0.

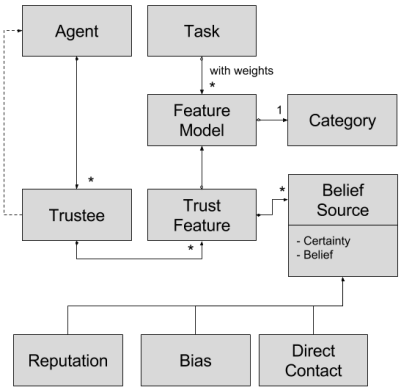

Figure 2: Memory Architecture (represented in UML)

4.1.1 Trust Calculation

Taking a Trustor X, a Trustee Y and a delegated task τ, Trust can then be calculated by taking the Trustee's Trust Features F_y, the Task's Feature Models F_τ and checking which they have in common, which we can represent as $F_{y\cap\tau}$. Remember that Trust Features are uniquely identified by a Feature Model. So after getting $F_{y\cap\tau}$ we can apply a linear function to each of the features in $F_{y\cap\tau}$, where for each element F_i we multiply the trustee's feature's belief value $B(F_i)$ with the weight of the feature for the task $W(F_i)$, as represented in Equation 6.

$$Trust_{X,Y,\tau} = \sum_{i=0}^{n} W(F_i)B(F_i) \tag{6}$$

The belief value of the feature itself, $B(F_i)$, is also calculated through a sum of parameters pertaining to each of the n belief sources $B^j_{F_i}$ composing the feature, as represented in Equation 7, with each parameter described as follows:

$$B(F_i) = \sum_{j=0}^{n} D^j_{F_i} C^j_{F_i} B_j \tag{7}$$

- $D^j_{F_i}$, a value from 0.0 to 1.0 that represents how far ago in time was this belief source received compared to the last one, being 0.0 a long time ago, and 1.0 the most recent belief. We wished to represent the rapid decay of value of old beliefs when compared to new ones, but also making sure recent memories would not fall quickly in value, so we chose to describe this parameter with a Gaussian Function, as represented in Equation 8, where $T^{Last}_{F_i}$ is the most recent belief value's time stamp, $T^j_{F_i}$ is $B^j_{F_i}$ belief value's time stamp, and L is the difference between the oldest and newest belief value's time stamps. We decided that $\frac{L}{4}$ defines a good mid drop-off point for the function.

- $C^j_{F_i}$, the certainty value stored in the Belief Source;

- $B^j_{F_i}$, the belief value stored in the Belief Source;

$$D^j_{F_i} = e^{-\frac{T^{Last}_{F_i} - T^j_{F_i}}{2(\frac{L}{4})^2}} \tag{8}$$

4.2 Perception

Another issue we encountered in literature was a lack of detail on how changes in the environment would be inserted into the model, so we try to solve that issue by defining relevant perceptions as part of the model. As a result, a variety of environment perceptions are defined in the model. This is done through a Perception object, representing some possible environment input, and containing a map of what target features should have belief sources added, what kind of belief sources they are, and how to translate the values received from the environment to belief value and certainty, as exemplified in Figure 3. When adding a Belief Source to a Trustee, if the associated Feature is not present, then it is added with the Belief Source. The affected Trustor and Trustee are received as arguments.

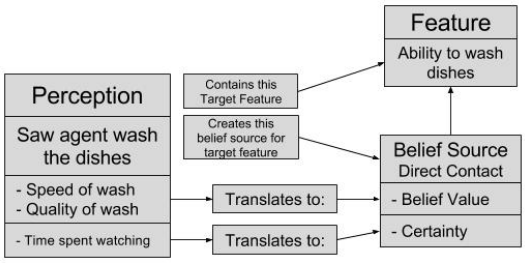

Figure 3: Perception Example

4.3 Action Suggestion

This is the module that is responsible for suggesting actions to the agent, in order to improve trust. It is composed by a series of Action objects, each represented by A and containing the following:

- $A_F = \{F_1, F_2, ..., F_i, ..., F_n\}$: A collection of n relevant Feature Models that this Action will affect. At least 1 Feature Model needs to be present in the action, but n is only limited by number available in the Model;

- $A_B(F_i) = \{B^{F_i}_1, B^{F_i}_2, ..., B^{F_i}_j, ..., B^{F_i}_m\}$: Each F_i Feature Model belonging to F has a collection of Belief Sources that describe how will the Feature be affected by the Action. Through this Belief Sources it is possible to predict how will the model change with this action, by inserting this Belief Sources in a mock model;

- $A_E = \{E_1, E_2, ..., E_i, ...E_p\}$: Each Action is mapped into p Environment Inputs, serving as flags to signal when it is appropriate to perform said Action;

- A_a: The action plan that is actually sent to the agent's planner. The definition of this plan is obviously dependent on the agent's architecture receiving the plan. While the complexity of this plans can achieve the implementation of social strategies, in the scope of this thesis, the actions were restricted to utterances that try to justify low ability or willingness (e.g. Saying that last game's low score was due to bad luck, but Jon can confirm my ability).

The Action Suggestion module tries to provide a suggestion when it is triggered by the reception of an Environment Input E_i. It then selects the Actions that have the received Environment Input mapped to them $E_i \in A_E$. The selected Actions are then sorted by a function S_F representing the potential increase in trust on the associated Features A_F. How S_F is defined is left as a parameter of the model, but we suggest a linear sum of all the differences in the affected features, as a simple solution (represented in Equation 9). After sorting through the selected Actions, the top-most ranked is sent to the Agent's Planner, and if it is in fact performed, the predicted Feature's Belief Sources are inserted into Memory. This process is represented in Figure 4.

$$S_F = \sum_{i=0}^{n} \Delta B(F_i) \tag{9}$$

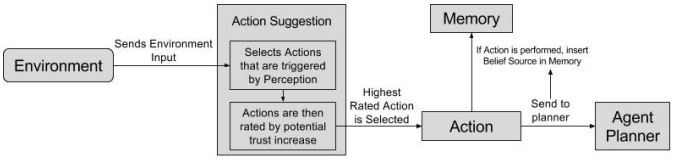

Figure 4: Action Suggestion Behaviour Flow

5. USER STUDIES

In order to evaluate the model we performed a user study in a Trust Game type scenario, with the intent to create an environment where the human participant would need to make a quantitative choice representing his trust in the agent.

5.1 Quick Numbers Scenario

As we approached the problem of evaluating the Trust Model proposed in this dissertation, we found that there was a lack of dedicated Trust evaluation scenarios that involved negotiation. Even in Game Theory based scenarios, we observed that there was a lack of attempts to encompass more than one dimension of trust. While the recent study by Salem, et. al. [31] addresses the role of robot task performance in trust, no study was found addressing perceived agent willingness to perform the task and its effect on trust.

While we were seeking for a solution, Henriques faced a similar problem in his thesis work on *Rapport - Establishing Harmonious Relationship Between Robots and Humans* [15], as he found no studies on robotic agents attempting to build Rapport using it's three components: positivity, coordination and mutual attention. Trust and Rapport are two very interconnected topics, with Rapport often seen as a strategy to increase trust. Due to this similarities, the overall scenarios that cover Trust also encompass Rapport analysis, so in an effort to better our respective evaluation phases we collaborated with Henriques to create a novel scenario: **Quick Numbers**. Based on the Trust Game [5], the scenario needs to be able to evaluate how both task performance and willingness jointly affect trust and observe all three components of rapport. The scenario was developed with the intention of evaluating a Trust model and a Rapport model, either separately or together. But Rapport related topics will not be mentioned in this paper, please refer to Henriques' thesis.

5.1.1 Scenario Overview

In Quick Numbers, a single human participant and a virtual agent are tasked to gain as many resources as possible. They both start with a fixed amount and are given the opportunity to multiply their resources by playing a simple eye-coordination game (further described in Section 5.2). The game starts by asking for a resource investment, and at the end, this investment is multiplied by an amount according to the player's performance and then given back to the player. The human and agent's games are independent from each other, but they are played at the same time and in opposite sides of a shared touch-screen table, so the human can socially interact with the agent and be able to perceive the agent's ability in the game. After both finish running through the game, the human will be asked to perform some task away from the agent. At this moment the virtual agent will give the participant the opportunity to invest in the agent's next game, but the participant is informed that the value given back to him is decided by the agent. In this phase, the virtual agent will have the opportunity to try and convince the human to invest or increase the investment by trying to manipulate trust. When the human returns the agent gives back as much as it wishes to give. This conjunction of different phases enables trust to be addressed in three distinct contexts: the ability to perform the task, willingness to perform the task, and willingness to return the investment.

5.2 Quick Numbers Game

While basing the scenario in the Investor Game, we decided that how the investee effectively multiplies the resources should be done by a task that the investor is at least familiar with. To this end, we have created a simple game concept consisting in a 2d rhythm game where the player must press numbered buttons in an increasing order (Figure 5), that spawn randomly in the screen, and disappear if not pressed after some time.

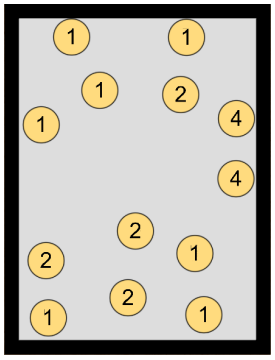

Figure 5: Quick Numbers Game

5.2.1 Agent's AI

An AI was created to progress through the scenario, with it being mostly scripted reactionary events where the Agent would just press a button when perceiving some specific change in the scenario. But one part we should discuss is the AI for the scenario's game, where although simple, some effort was given to make it parametrizable, in order to adjust the agent's ability in the game. The AI was programmed to press one of the available circles in a timed cycle, with 3 parameters:

- *Clicking Interval* (C_i): the amount of time between presses in seconds, so one of the circles is pressed every C_i seconds;
- *Chance of Right Target* (C_r): when pressing one of the circles, the AI will choose what circle to press, the chance that it will choose the correct one is given by C_r;
- *Reaction Time* (R_t): a circle is only be eligible to be pressed by the AI R_t seconds after it spawns, as to replicate the reaction time a human would have to recognize the circle.

We wanted the agent to play rapidly, if a little recklessly, as to provide more opportunities for the participant to react

and notice how he played, so we empirically found 0.5 seconds to be a good value for C_i, as it made the agent have a slightly above average human speed to press the circles. To C_r we assigned 70% success rate, as failing 30% of the circles averaged out the agent's score to normal human achievable levels, and to R_t we selected 0.3 seconds, as it is a plausible value to accompany a 30% fail chance, as it simulates the agent not entirely recognizing the number before pressing the circle.

6. AGENT ARCHITECTURE

The agent we used to serve as a host to the Trust Model was built using Henriques' Rapport Controller [15], a computational framework developed to create human agent interactions and transmit them to the various components that control the agent embodiment. The Rapport Controller uses a plug-in architecture, with most of the behaviour inserted through this plug-ins.

6.1 Trust Model Plug-in

The Trust Model was implemented as a plug-in to the Rapport Controller. Although most of the model is implemented as described in Section 4, there were some simplifications made to the model:

- Trust Calculation described in Section 4.1.1 does not include the $D^j_{F_i}$ parameter, as the amount of time that passes in the scenario is negligible;
- Action Suggestion lacks Action sorting and selection as described in Section 4.3. Instead only one action is ever associated to an Environment Input, but performing only if the current Trust Value is lower than constant associated to the Action;
- Perceptions do not receive the agents as input. Each Perception contains the Trustor and Trustee agents affected in the interaction perceived;
- Certainty values were inserted at the identity value of 1.0f, therefore having no effect in model calculation.

7. METHODOLOGY AND PROCEDURES

The study was conducted with a between subject design with the following conditions:

- **Condition B**: a baseline condition, where the Action Suggestion component is not active. The data gathered in this condition will serve as the basis to which we compare our results;
- **Condition T**: the condition where Action Suggestion is active, serving as the main results condition.

The user study sessions were individual and performed in an closed room accompanied just by the researcher, and lasted between 20 and 30 minutes. The sessions followed the scenario overview as described in Section 5.1.1, with the interactions performed through the agent and a touch-screen table (Figure 6). Additionally the participants answered a questionnaire, divided in 3 parts, to be filled in different stages of the scenario:

Figure 6: Participant playing with EMYS.

- The first part gathers demographic and sampling information, like gender, age and if the participant had previous interactions with EMotive headY System (EMYS). It also evaluates a participant's self-trust and inclination to trust in others, through a series of questions created by Carrington [7]. Participants were asked to fill this before starting the scenario;
- The second part is composed of a simple question to self-evaluate trust in the agent, and the trust perception scale by Schaefer [33]. This Section was asked to filled at the end of the Investment Stage, immediately after the participant invested on EMYS. In fact this serves as the other task that the participant must be doing while the agent plays the game alone;
- The third and final part is a GodSpeed questionnaire [4, 20] and a proximity scale [3]. It is answered by the participants at the end of the scenario.

8. RESULTS

To evaluate if Trust was improved by the introduction of our Action Suggestion module, we used two Trust measurements: one obtained in the questionnaire, in the Schaefer section, and the other through the Investment value retrieved from the scenario. Then we compared the results between conditions B and T. Therefore, the following hypothesis for the study arose:

- Are Trust levels improved by the inclusion of the Action Suggestion module?
- Does the participant's Investment value in the scenario increase by the inclusion of the Action Suggestion module?

All statistical analyses further mentioned used a significance level of 5%.

Are Trust levels improved by the inclusion of the Action Suggestion module?

To infer a conclusion on this hypothesis we compare the means of the results obtained from the Schaefer section questionnaires, using the Independent-Samples T-Test to check their significance. A Shapiro-Wilk normality test was also performed to conform to the T-Test sample normality assumption. As seen in the Box-plot represented in Figure 7 there is no significant apparent differences between results in condition B and T, further confirmed by checking the very low difference between means represented in the measurement descriptives in Table 1, supported by a very high

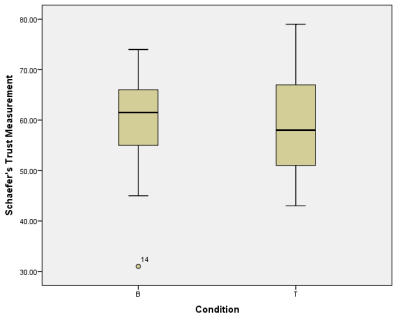

Figure 7: Box-plot of Schaefer measurement results (Condition B Median: 61.5; Condition T Median: 58.0).

Descriptives	Condition B	Condition T
Mean	59.05±2.32	59.47±2.64
Std. Deviation	10.38	10.90
Shapiro-Wilk Sig.	0.157	0.622
T-Test Mean Difference	-0.421	0.421
T-Test Sig.	0.905	0.906

Table 1: Schaefer Measurements Descriptives.

significance value in T-Test, inferring no significant difference between the means.

Answer: There were no significant differences between the means of Schaefer's Trust measurements in the 2 conditions.

Does the participant's Investment value in the scenario increase by the inclusion of the Action Suggestion module?

Due to the distribution of the Investment value not being normal in condition B, as observed in the histograms in Figures 8 and 9, we used Mann–Whitney U statistical test to determine if there is a significant difference between the results in each condition. Additionally, as the distribution shapes are quite different, we can only check through mean rank values. But with a significance p-value of $p = 0.707$ in the Mann–Whitney U test we cannot conclude any significant difference in results between the conditions, evidenced further on the box-plot graph represented in Figure 10.

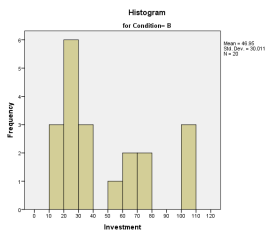

Figure 8: Investment values in Condition B Histogram

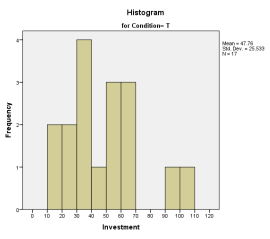

Figure 9: Investment values in Condition T Histogram

Answer: There were no significant differences between the means of Investment value measurements in the 2 conditions.

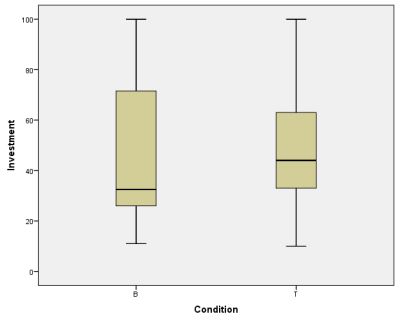

Figure 10: Box-plot of scenario Investment measurement results (Condition B Median: 32.5; Condition T Median: 44.00).

8.1 Results Discussion

The study results show no statistical significant change in trust measurements between conditions B and T, leading to inconclusive results. But by observation of the box-plot graphs, seen in Figures 7 and 10, it seems that the Action Suggestion module had no effect on the participant's trust in the agent. We believe that this results are due to the oversimplification of the implemented model, not only were the Actions just utterances, but these utterances were not properly verified by experts as appropriate to increase Trust. The number of actions was also very few, leading to a lack of agency. Additionally, the participants commented that they could not pay that much attention to how the agent played, as the game required too much attention, so the games should be played non-concurrently, in order to give the participant opportunity to focus on how the agent plays the game.

9. CONCLUSIONS

Throughout this paper we addressed the work done to develop a Cognitive Trust Model capable of suggesting actions to improve Trust in a virtual agent. We first went through our thoughts about the lack of research done in the area of trust in HRI, especially regarding trust improvement, and what we propose to address that issue. Then we went on to establish some background in HRI concepts specific to the domain. In the next chapter we delved into some of the Cognitive Trust Models that we found. While there were many more aside from those discussed in this thesis, we wanted to focus on the ones that most closely related with the one we developed. Following that we presented our Trust Model, describing its 3 main components: Memory, Perception and Action Suggestion, and how they interact to compose our model. Due to problems finding a suitable evaluation scenario for our model, we then describe our second contribution of this thesis, a novel Trust and Rapport evaluation scenario, Quick Numbers, that aims to evaluate how ability and willingness jointly affect trust, by imposing to the participant the choice to entrust the agent, with their own earned resources, to play a game, in which the participant has some idea of the agent's ability. Finally we showed how we used the scenario to perform User Studies on the Trust Model. Unfortunately, the results can be considered either inconclusive, or right against our effort to improve Trust, as there was no apparent change in Trust measurements in

conditions with and without the Action Suggestion module. Nevertheless, we believe that our contributions were significant by providing an implementable base from which other research projects in the same area may start.

10. REFERENCES

[1] A. Abdul-rahman and S. Hailes. Supporting Trust in Virtual Communities. *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*, 00(c):1–9, 2000.

[2] J. Allen, N. Chambers, G. Ferguson, L. Galescu, H. Jung, and W. Taysom. PLOW : A Collaborative Task Learning Agent. *Interpreting*, 22:1514–1519, 2007.

[3] A. Aron, E. N. Aron, and D. Smollan. Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, 63(4):596–612, 1992.

[4] C. Bartneck, D. Kulić, E. Croft, and S. Zoghbi. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics*, 1(1):71–81, jan 2009.

[5] J. Berg, J. Dickhaut, and K. McCabe. Trust, Reciprocity, and Social History. *Games and Economic Behavior*, 10(1):122–142, 1995.

[6] T. W. Bickmore and R. W. Picard. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Transactions on Computer-Human*, 12(2):293–327, 2005.

[7] K. Carrington. Toward the development of a new multidimensional trust scale. (March):1–366, 2007.

[8] C. Castelfranchi and R. Falcone. Principles of trust for MAS: cognitive anatomy, social importance, and quantification. *Proceedings of the International Conference on Multi Agent Systems*, pages 72–79, 1998.

[9] C. Castelfranchi and R. Falcone. *Trust Theory*. John Wiley & Sons, Ltd, Chichester, UK, 1 edition, mar 2010.

[10] D. Gambetta. Can We Trust Trust? In *Trust: Making and Breaking Cooperative Relations*, pages 213–237. Blackwell, 1988.

[11] M. a. Goodrich and A. C. Schultz. Human-Robot Interaction: A Survey. *Foundations and Trends® in Human-Computer Interaction*, 1(3):203–275, 2007.

[12] J. Granatyr, V. Botelho, O. R. Lessing, E. E. Scalabrin, J.-P. Barthès, and F. Enembreck. Trust and Reputation Models for Multiagent Systems. *ACM Computing Surveys*, 48(2):1–42, oct 2015.

[13] B. J. Grosz. Collaborative Systems. *AI Magazine*, pages 67–85, 1996.

[14] Han Yu, Zhiqi Shen, C. Leung, Chunyan Miao, and V. R. Lesser. A Survey of Multi-Agent Trust Management Systems. *IEEE Access*, 1:35–50, 2013.

[15] B. Henriques. Rapport - Establishing Harmonious Relationship Between Robots and Humans, 2016.

[16] H. Huang, G. Zhu, and S. Jin. Revisiting Trust and Reputation in Multi-agent Systems. *Computing, Communication, Control, and Management, 2008. CCCM '08. ISECS International Colloquium on*, 1:424–429, 2008.

[17] T. D. Huynh, N. R. Jennings, and N. R. Shadbolt. An integrated trust and reputation model for open multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 13(2):119–154, 2006.

[18] S. Jones and S. Marsh. Human-computer-human interaction. *ACM SIGCHI Bulletin*, 29(3):36–40, jul 1997.

[19] J. D. Lee and K. A. See. Trust in Automation : Designing for Appropriate Reliance. 46(1):50–80, 2004.

[20] H. Lehmann, J. Saez-Pons, D. S. Syrdal, and K. Dautenhahn. In Good Company? Perception of Movement Synchrony of a Non-Anthropomorphic Robot. *PLOS ONE*, 10(5):e0127747, may 2015.

[21] S. P. Marsh. *Formalising Trust as a Computational Concept*. PhD thesis, apr 1994.

[22] Z. Noorian and M. Ulieru. The State of the Art in Trust and Reputation Systems: A Framework for Comparison. *Journal of theoretical and applied electronic commerce research*, 5(2):97–117, aug 2010.

[23] I. Pinyol. Reputation-Based Decisions for Cognitive Agents (Thesis Abstract). *Doctoral Mentoring Program*, (Aamas):33, 2009.

[24] I. Pinyol and J. Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 40(1):1–25, jun 2013.

[25] A. S. Rao and M. P. Georgeff. BDI agents: From theory to practice. *Icmas*, 95:312–319, 1995.

[26] D. Rousseau, S. Sitkin, R. Burt, and C. Camerer. Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, 23(3):393–404, 1998.

[27] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach, 3rd edition*. 2009.

[28] J. Sabater, M. Paolucci, and R. Conte. Repage: REPutation and ImAGE among limited autonomous partners. *Jasss*, 9(2):117–134, 2006.

[29] J. Sabater and C. Sierra. Reputation and social network analysis in multi-agent systems. *Proceedings of the first international joint conference on Autonomous agents and multiagent systems part 1 - AAMAS '02*, page 475, 2002.

[30] J. Sabater and C. Sierra. Review on computational trust and reputation models. *Artificial Intelligence Review*, 24(1):33–60, 2005.

[31] M. Salem, G. Lakatos, F. Amirabdollahian, and K. Dautenhahn. Would You Trust a (Faulty) Robot? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*, pages 141–148, New York, New York, USA, 2015. ACM Press.

[32] K. Schaefer. *The Perception and Measurement of Human-Robot Trust*. PhD thesis, 2009.

[33] K. E. Schaefer. The perception and measurment of human-robot trust. 2013.

[34] J. A. Simpson. Foundations of interpersonal trust. In *Social psychology: Handbook of basic principles (2nd ed.)*., pages 587–607. 2007.

[35] R. van den Brule, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, and W. F. G. Haselager. Do Robot Performance and Behavioral Style affect Human Trust ? *International Journal of Social Robotics*, 2014.