Trustful Action Suggestion in Human Agent Interaction

[Extended Abstract]

Nuno Xu
Instituto Superior Técnico
1932 Wallamaloo Lane
Wallamaloo, New Zealand
nuno.xu@tecnico.ulisboa.pt

ABSTRACT

This paper provides a sample of a LaTeX document which conforms, somewhat loosely, to the formatting guidelines for ACM SIG Proceedings. It is an *alternate* style which produces a *tighter-looking* paper and was designed in response to concerns expressed, by authors, over page-budgets. It complements the document *Author's (Alternate) Guide to Preparing ACM SIG Proceedings Using LaTeX2$_\epsilon$ and BibTeX*. This source file has been written with the intention of being compiled under LaTeX2$_\epsilon$ and BibTeX.

The developers have tried to include every imaginable sort of "bells and whistles", such as a subtitle, footnotes on title, subtitle and authors, as well as in the text, and every optional component (e.g. Acknowledgments, Additional Authors, Appendices), not to mention examples of equations, theorems, tables and figures.

To make best use of this sample document, run it through LaTeX and BibTeX, and compare this source code with the printed output produced by the dvi file. A compiled PDF version is available on the web page to help you with the 'look and feel'.

Keywords

ACM proceedings; LaTeX; text tagging

1. INTRODUCTION

Trust has been described in Psychology as being one of the most important components of interpersonal relationships [30]. 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 Artificial Intelligence (AI) research gravitates towards the development of Intelligent Agent Systems [24], a focal concern is the performance of collaborative tasks[11, 3, 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[5] and the sub-field of Human-Robot Interaction (HRI)[9, 31], specially when so much has been done regarding Trust in Automation (TiA) [16, 15, 17]. This reveals that while the area has so much potential, the level of understanding is still very shallow, only deeply focused in certain areas[10].

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 the internal trust model structure of the agent, normally with just the purpose of ranking other agents, through the use of statistical and game theoretical based methods[10]. This makes it difficult to create a model that is easy to understand, analyse and, most importantly, describe its evaluative reasoning in a human understandable manner. The introduction of cognitive models by Castelfranchi and Falcone [6] tries to solve that problem by mapping the trust model to the agent's mental state, composed by beliefs and goals, very akin to existing cognitive agent architectures like BDI[22]. Then some systems, like Repage[25], created implementations of this new paradigm of trust modelling; until then 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 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 a greater degree of integration with the trust model of the agent. To our knowledge, no attempts have been done towards this goal, so we propose to develop two agent modules: firstly, one capable of creating a cognitive model representing the mental state of the user's trust in the agent, using Repage's architecture, and secondly, another to suggest what actions should be used to improve trust on the agent. We will ascertain this project's objectives by integrating the modules in an agent implementation that is capable of acting as one of the players in the *Split or Steal* scenario, introduced in the British

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

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

game show *Golden Balls*[2] (currently finishing development in our research group, GAIPS[3]). The scenario is further described in Section **??**.

We hope that this project will make agent decision making more interesting, provide some insight on how actions affect trust and budge the field a bit in this unexplored direction.

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 report, 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[23, 15, 27]. 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[7]. In the scope of this project, 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 [30]), 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[8] 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 [7]). 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[18] 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[8], 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 [7]). 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[6]. 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[6] 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'[7].

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[6].

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, 26, 25, 14, 20]), where more recent trust models have been developed to also include reputation as a source of trust.

An agent is not influenced only 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 as introduced by Sabater in [25]: 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 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.

[2] Golden Balls TV Show: http://www.goldenballstvshow.com/

[3] Intelligent Agents and Synthetic Characters Group (GAIPS): http://gaips.inesc-id.pt/

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[10, 12, 21, 19, 13], with at least 106 models created[10], since the formalization of trust as a measurable property by Marsh in 1994 [18]. 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[6], 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 want to focus 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 [18].

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 will not use the entirety of this model, it is worth describing, as it was also a source of inspiration to other authors referenced in this report. 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.

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

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.

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

$F_{X,Y,\tau}$ 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 for our modules such adaptability is interesting for our intent to try our modules in different scenarios.

3.1.2 Repage: A REPutation and ImAGE model

This system was introduced in 2006 by Sabater et al.[25] 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* (check Figure 1). 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.

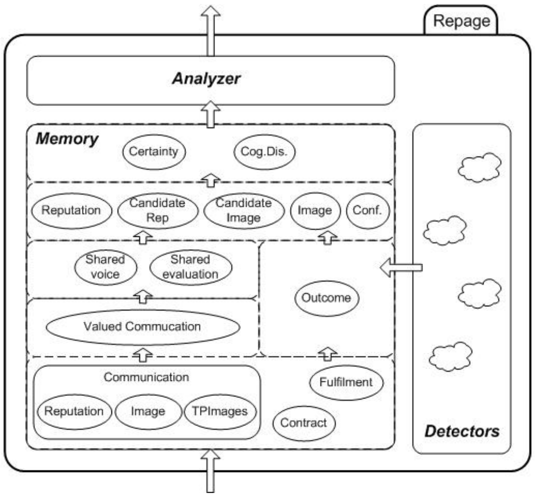

Figure 1: Repage architecture schematic (taken from [25])

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.

3.1.3 Discussion

Of the related work discussed here, we are going to base our solution on Repage and *BC*-logic, as described in their respective Sections 3.1.2 and **??**. Repage fits well as a basis for our objectives, as it has the details of modelling trust already dealt with and leaves us the room to develop the analysis component that corresponds directly to the goal of this project. The choice was also made out of convenience, as no other work was found were implementable design was a concern.

3.2 The Perception and Measurement of Human-Robot Trust

Schaefer[29] presents a trust perception scale providing a way of extracting an accurate trust score from humans interacting with robots. The scale is composed of 40 items that can be ranked from 0 to 100, in 10 point intervals. The final result it then averaged by adding all the item values and divided by the total number of items (40).

While this work has been done specifically for HRI we believe that a sub-set of this items can be used for the features used in the cognitive model of the user's trust, further described in Section **??**. The items are listed in Table **??** in appendix **??**.

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.[25] 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 participating in a social interaction, 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 agent's in the scenario, and the model created would be his own trust model of the world environment. And so, the model can be described by 3 main components:

- **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.

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 [7] as a basis for trust definition, 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 ($\boldsymbol{gx}$) 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 the known entities in the scenario world space, serving mostly as an identifier;
- **Trustee**: each agent contains a collection of other agents he 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 represents 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;
- **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 provided 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.

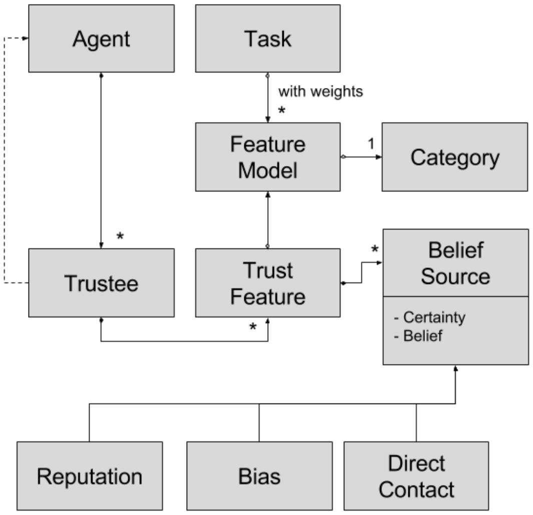

Figure 2: Memory Architecture (represented in UML)

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. $\frac{L}{4}$ defines the mid drop-off point of 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 would changes 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.

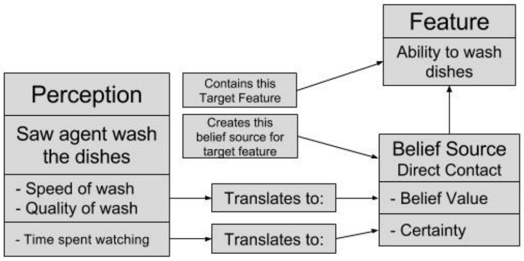

Figure 3: Perception Example

4.3 Action Suggestion

This module contains a series of Actions that are suggested to the agent's planner when certain conditions are met. This Actions should be able to encompass a variety of social strategies relevant to the scenario, but we reduced the scope in this thesis to just utterances suggestions. Each Action is associated to the Features that it predicts it is going to change, storing a Belief Source for each Feature. If a suggested Action is performed it's Belief Sources are then added to the model. A suggestion occurs when an environment perception is received and the current world state is classified as a good moment to perform one of the stored Actions, of course not all Actions will be relevant in a particular situation, so only some of them are then sorted by importance, by checking which action will positively affect the features with the lowest belief values.

5. USER STUDIES

In order to evaluate the model we performed a user study in a Investor 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. These studies were performed in collaboration with Henriques for his thesis work on *Rapport - Establishing Harmonious Relationship Between Robots and Humans* [**?**], as by having similar evaluation goals we decided that a joint effort would improve the overall quality of our work.

5.1 Quick Numbers Scenario

As we approached the problem of evaluating the models proposed in this dissertation, we found that there was a lack of dedicated Trust or Rapport evaluation scenarios. On one hand, there was a lack of attempts to encompass more than one dimension of trust, and on the other, there was a lack of studies that approach Rapport on robotic agents while at the same time, attempt to build rapport using the three components of rapport: positivity, coordination and mutual attention. While the recent study by Salem, et. al.[28] 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. Therefore, we created a novel scenario *Quick Numbers*, based on the Trust Game [4], that would enable us to evaluate how both task performance and willingness would 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.

In *Quick Numbers*, a human subject is tasked with gaining as many resources as possible. He starts with a fixed amount and is given the opportunity of playing a game that can multiply his resources. The game starts by asking for a resource investment, and at the end, a multiplied amount of this investment is given back according to the player's performance. The virtual agent will also participate in this first step as though it is also a subject, also playing the game simultaneously with the human and sharing the same goal of getting the maximum amount of resources. At this point, the human should be able to perceive the agent's ability in the game. After finishing the game, the human will be asked to fill out a questionnaire, and the virtual agent will give the human the opportunity to invest in the next game, increasing the maximum potential return. 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.

In order to evaluate rapport and trust, participants will answer 3 questionnaires, one before playing with the virtual agent, one after investing on the agent but before knowing the result, and another after the investment return is given back. In the following paragraphs, we will detail the scenario, the procedures and methodologies of the study applied to our thesis, followed by the sample description.

5.1.1 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 reaction time for pressing 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 plausible value for a 30% fail chance, as the agent simulates not entirely recognizing the number before pressing the circle.

5.2 Trust Game

Our overall scenario is very much based on the Trust/Investor Game, first proposed by Berg et al. [4], so we present a small introduction and discussion to the scenario. The game is set up with 2 anonymous players, which we will call player A and player B, where $10 is given to player A and none to player B. In the first phase player A must choose how much of the starting $10 should he give to player B knowing that the value will be tripled in player B's hands. In the second phase player B chooses how much of the, now tripled, money will he return no player A.

This game forms a good base for our scenario because the decision of how much A should give to B is dependent on 2 different factors: the ability and willingness of B to multiply the investment and the willingness of B to return the profits of the investment. This is possible to perform because, while the source of the game explicit the multiplication factor of giving money to B, this value can be a hidden parameter, making the ability of B something to take into account.

We used this game as a foundation for our user studies, described further on in Section 5, where we attempted to resolve some of it's faults, such as the game lack of any negotiation phase, in fact, both players are in separate rooms, with no way of interacting with one another, inducing trust to be modelled in a game theory point of view, which is contrary what we wish to accomplish.

5.3 Results

6. CONCLUSIONS

This paragraph will end the body of this sample document. Remember that you might still have Acknowledgments or Appendices; brief samples of these follow. There is still the Bibliography to deal with; and we will make a disclaimer about that here: with the exception of the reference to the LaTeX book, the citations in this paper are to articles which have nothing to do with the present subject and are used as examples only.

7. ACKNOWLEDGMENTS

This section is optional; it is a location for you to acknowledge grants, funding, editing assistance and what have you. In the present case, for example, the authors would like to thank Gerald Murray of ACM for his help in codifying this *Author's Guide* and the **.cls** and **.tex** files that it describes.

8. 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] J. Allen and G. Ferguson. Human-machine collaborative planning. *International NASA Workshop on Planning*, pages 1–10, 2002.

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

[5] 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.

[6] 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.

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

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

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] J. Lee and N. Moray. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10):1243–70, oct 1992.

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

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

[19] 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.

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

[21] 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.

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

[23] 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.

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

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

[26] 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.

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

[28] 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.

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

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

[31] 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.

APPENDIX

A. HEADINGS IN APPENDICES

The rules about hierarchical headings discussed above for the body of the article are different in the appendices. In the **appendix** environment, the command **section** is used to indicate the start of each Appendix, with alphabetic order designation (i.e. the first is A, the second B, etc.) and a title (if you include one). So, if you need hierarchical structure *within* an Appendix, start with **subsection** as the highest level. Here is an outline of the body of this document in Appendix-appropriate form:

A.1 Introduction

B. MORE HELP FOR THE HARDY

The sig-alternate.cls file itself is chock-full of succinct and helpful comments. If you consider yourself a moderately experienced to expert user of LaTeX, you may find reading it useful but please remember not to change it.