Trust in Human and Autonomous Agent Interaction (Maybe change to: Trustful Action Suggestion in Human Agent Interaction)

Nuno Xu
nuno.xu@tecnico.ulisboa.pt

Supervisor: Rui Prada
Co-Supervisor: Ana Paiva
Técnico Lisboa (Taguspark)
Universidade de Lisboa
Av. Prof. Dr. Aníbal Cavaco Silva
Porto Salvo, Portugal
http://tecnico.ulisboa.pt/en/

Abstract What thesis is about and contributions

Key words: rapport, Human Robot Interaction (HRI)

Contents

1 Introduction

Trust has been described in Psychology as being one of the most important components of interpersonal relationships [1]. It is undeniable the need of Trust to promote cooperation and collaboration between two parties, either in deciding who should one collaborate with, or even on what exactly do we trust the other party with.

As Artificial Intelligence (AI) Research gravitates towards the development of Intelligent Agent Systems [2], out of which a focal concern is the performance of collaborative tasks [3–5], as well as addressing the problems of interaction between humans and agents [6], one would consider that Trust should be one of the main focuses of Human Agent Interaction (HAI). Since the start of automated machinery, one of the main issues was how to properly manage trust on machines, in order to avoid over or under reliance [7]. Reeves and Nass have shown that people apply social rules to Human Computer Interaction (HCI), and this can logically be extended to the sub-field of HAI [8]. So as agents evolve to better perform collaborative tasks with humans autonomously, which demands at least some amount of social interaction, the active agent must seek out to improve the trust relationship it has with the user [9]. And while the amount of literature has been increasing, we found it surprising that not enough work has been done in HAI focusing on Trust, other than on design issues [10] and the sub-field of HRI [11, 12], specially when so much has been done regarding Trust in Automation [7, 13, 14]. This reveals that while the area has so much potential, the level of understanding is still very shallow, only deeply focused in specific areas.

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 *E-bay* [15], where 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. 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 [16] tries to solve that problem, mapping the trust model to the agent's mental state, composed by beliefs and goals, very akin to existing cognitive agent architectures like BDI [17]. Then some systems, like Repage [18], 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, and that is 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 (currently finishing development in our research group, GAIPS[1]) that is capable of acting as one of the players in the *Split or Steal* scenario, introduced in the British game show *Golden Balls* [19], the scenario is further described in Section 5.2.

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

1.1 Goals

In sum, this project's purpose will be to:

- Create a cognitive trust model capable of representing human trust towards the agent using the Repage architecture;
- Develop an action suggestion module that aims to provide actions that improve trust in the agent (or a least the beliefs on this trust) for the Repage;
- Calibrate the modules through user testing in the *Split or Steal* scenario.

In the remainder of the document we will present a brief summary of the main concepts used in this project in Section 2. Then in Section 3 we will discuss some of the work done in modelling trust for MASs and measuring trust in HRI applications. In Section 4 a description of our solution architecture will be presented, followed by our plans and schedules in Section 6. Finally in Section 5 we will describe how we will evaluate the project.

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

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

Sociology, to Philosophy and Economy [14, 20, 21]. 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 [22]. 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 [1]), 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 [23] 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 [22]). 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 [24] 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 [23], 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 [22]). 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 [16]. 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 [16] 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' [22].

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 [16].

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 [18, 25–28]), 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 [18]: 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.

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. To better explain the concepts we want to present, we will introduce one of the most common exemplary models of Game Theory, the Prisoner's Dilemma.

2.3.1 Prisoner's Dilemma The Prisoner's Dilemma is a two player game and is usually described as follows:

Two criminal partners are arrested and locked in separate cells with no way of communicating with each other. They are then questioned separately, where they are given 2 options, betray the other prisoner by testifying against him, or remain silent, with the following outcomes:

- If both prisoners betray each other, both get 2 years in prison;
- If one of them betrays and the other remains silent, the betrayer goes free and the other gets 3 years in prison;
- If both remain silent, both get just 1 year in prison;

We can represent betraying as *Defecting* (D), and staying silent as *Cooperating* (C), and name the players *player1* and *player2*. So the game's possible outcomes can be represented by a payoff matrix, like the one in Table 1 where each entry represents a tuple of the form (*player1* payoff, *player2* payoff). As the goal is to not get years in prison, the payoffs correspond to $Max\ years\ in\ prison - years\ got\ in\ prison$.

	C_2	D_2
C_1	2,2	0,3
D_1	3,0	1,1

Table 1: Prisoner's Dilemma Payoff Matrix

In the game we can say that *Defecting* **dominates** *Cooperating*, as for any action that the adversary player may choose, *Defecting* always gives a better payoff for the individual player [29].

3 Related Work

Computational Trust research has been focused on modelling trust in MASs, specially on open e-commerce environments [30–34], with at least 106 models created [30], since the formalization of trust as a measurable property by Marsh in 1994 [24]. 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 [16], 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 [24].

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.

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

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 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.* [18] 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 interconnected 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 thresh-

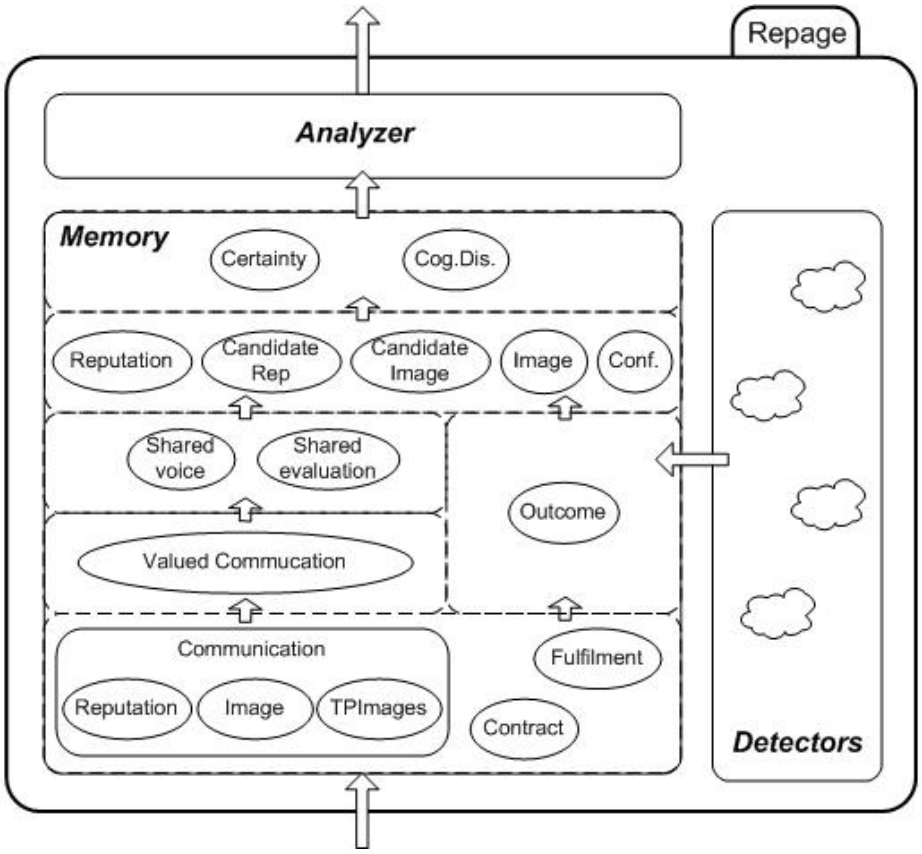

Figure 1: Repage architecture schematic (taken from [18])

old 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 *BC*-logic: a representation of beliefs for Repage Pinyol *et al.* [28] proposes an integration of the Repage model, introduced in the previous Section 3.1.2 with a BDI Agent [17]. While the BDI model is not relevant to us, their work specifies *BC*-logic, a belief first order logic that is capable of representing Repage predicate semantics and this is the part we will describe in the following paragraphs.

BC-logic is structured hierarchically, in a way that formulas from a certain first-order language lower in the hierarchy can be embedded in another language above as constants. This is written as $\lceil\phi\rceil$, with ϕ being the formula of the lower language. The hierarchy is composed of three languages, starting with the base language, L_{basic}, that expresses the ontology and contains the symbols to represent the domain. Next there is L_{ag}, which contains symbols of the base language and special predicates to allow to reason about probability of formulas, about formulas communicated, and formulas believed by agents. Finally there is *BC*-language, the meta-logic language, with the aim to express statements about the agents' reasoning. L_{ag} and *BC*-language are sorted languages, in other words, its symbols and predicates are partitioned into sorts, each containing their own semantics. All languages contain the logical symbols $\forall$, $\exists$, $\wedge$, $\vee$, $\neg$, and $\Longrightarrow$.

L_{ag} contains four sorts:

- S_D: represents application domain, including constants, functions and predicate symbols;
- S_R: represents probabilities, including a set of constants, C_R, with a label written as $\overline{r}$, where $r \in [0,1] \bigcap Q$;
- S_A: represents agent names, including a set of constants $C_A = i_1, ..., i_n$, corresponding to the agents' identifiers;
- S_F: represents formulas, including a set of constants C_F, which is built simultaneously with the construction of the language. This is done by adding the constant $\lceil\phi\rceil$ for each $\phi \in F_m(L_{basic})$, and then, given a formula $\Psi \in F_m(L_{ag})$ we also add $\lceil\Psi\rceil$ to C_f.

Symbols in predicates are identified by their sorts, take for example a binary predicate B, it is written as $B(S_A, S_F)$, meaning that the first argument must be part of sort S_A and the second argument part of sort S_F.

The set of formulas $Fm(L_{ag})$ has the following special predicates:

- $B(S_A, S_F)$: An agent's belief towards a formula (e.g. $B(i_c, \lceil sunny(Lisbon) \rceil)$); abbreviated to $B_{S_A}(S_F)$;
- $Pr \leq (S_F, S_R)$, $Pr \geq (C_F, C_R)$: A lower/upper bound on probability of a formula (e.g. $Pr \geq (\lceil sunny(Lisbon) \rceil, 0.8)$);
- $S(S_A, S_A, S_F)$: The communication predicate, as stated in Repage (e.g. $S(i_c, j_c, \lceil sunny(Lisbon) \rceil)$); abbreviated to $S_{S_A,S_A}(S_F)$.

BC-language contains five sorts:

- S_R, S_A and S_F: as defined above for L_{ag};
- S_V: represents variable sequences;
- S_T: represents ground term sequences;

Image and Reputation towards an agent j_c, playing the role r can then be represented by:

- Image: $Img_{i_c}(j_c, r, [V_{w_1}, ..., V_{w_m}])$
- Reputation: $Rep_{i_c}(j_c, r, [V_{w_1}, ..., V_{w_m}])$

with $[V_{w_1}, ..., V_{w_m}]$ being an abstracted set of evaluations in the belief, as while Repage maps evaluation from Very Bad to Very Good, this can be applied to any ordered mapping of m evaluations. As a simplification, the model summarizes the interaction model of the participating agents (i_c and j_c) to a single action. Through this a mapping $R_r i$ can be defined between each role r, agent i_c and the action. A mapping T_{r,w_k} is also defined between each role r and label w_k to a formula written in L_{basic}.

Image and Reputation can also be represented as a set of beliefs:

$$\begin{array}{cc} \underline{Img_{i_c}(j_c, r, [V_{w_1}, ..., V_{w_m}])} & \underline{Rep_{i_c}(j_c, r, [V_{w_1}, ..., V_{w_m}])} \\ B_{i_c}(pr \geq ([R_{rj_c} T_{r_{w_1}}, V_{w_1}])) & B_{i_c}(s(pr \geq ([R_{rj_c} T_{r_{w_1}}, V_{w_1}]))) \\ B_{i_c}(pr \geq ([R_{rj_c} T_{r_{w_2}}, V_{w_2}])) & B_{i_c}(s(pr \geq ([R_{rj_c} T_{r_{w_2}}, V_{w_2}]))) \\ ... & ... \end{array}$$

The work goes into further detail regarding representing the relationship of Image and Reputation, addressing agent honesty and consistency in communicating reputation, and while interesting, it is not part of what we to model.

Overall *BC*-logic is an interesting approach to representing beliefs in the Repage model and we will most likely choose it for the model representation, as it is the most well developed that we found.

3.1.4 Sutcliffe and Wang's model This work was published by Sucliffe and Wang in 2012 [35] and they built a trust model to figure out how cognitive social mechanisms emerge to follow Dunbar's Social Brain Hypothesis (SBH) [**?**], an evolutionary psychology theory that proposes that humans have a predisposition

to build relationships in layers of decreasing intimacy. As trust has been acknowledged to be one of major component of human relationships, they demonstrate that simulating trust development and decay, through interactions and neglect, respectively, show the patterns predicted in SBH.

From the model standpoint, its main interesting feature is that agents develop trusting relationships between one another, affecting interaction frequency between agents, by preferring to pick those with already high trust value. Additionally the trust relationship degrades as time passes by, with variable speeds depending on the current relationship level, giving stronger ties a slower descent. All in all, the model provides a good tool to simulate multi agent social behaviour, and may be interesting to predict trust degradation in the agent, albeit its described application for social simulations is a bit far from our scope.

3.1.5 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 3.1.3. 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 [36] 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 4.1. The items are listed in Table 3 in appendix A.

4 Proposed Solution

In this Section we will address the components we will develop, capable of suggesting trustful actions to a generic autonomous agent. Our solution will be using the Repage architecture, described in Section 3.1.2, as it is the only implementable trust modelling architecture, that we found, created to be implementable, and not just a theoretical model. It's simplicity also goes in line with our goals of creating an easily comprehensible model, and developing a complete trust architecture is outside the scope of this project.

We will start with describing a trust model of the user in the Cognitive Trust Modelling Module (Section 4.2), go on to the Trustful Action Suggestion Module (Section 4.2) that will actually handle action suggestion, and finally talk about how they connect together and to the rest of a generic agent architecture.

4.1 Cognitive Trust Modelling Module

In this module we aim to create a trust representation of the user. The model must be able to represent the user's trust beliefs, while also provide an evaluation on how trustworthy is the agent in the eyes of the user.

This module will represent the memory and detector components of the Repage architecture (Section 3.1.2). The concrete implementation for the beliefs is still under discussion, but our main candidate would be the *BC*-Logic described in 3.1.3 by Pinyol *et al.* [28].

4.2 Trustful Action Suggestion Module

This module is still very roughly defined as no related work was found about this topic, so most work done in this module will be on experimenting what information can be extracted from the trust model. The main goal is to suggest actions that will either improve the strength of existing beliefs on the trust model, or improve the trust value on the agent, occupying the analyser component in Repage (Section 3.1.2).

4.3 Putting it Together

The Trustful Action Suggestion Module is going to be directly dependent on the Cognitive Trust Modelling Module, as all decision making from the former module is going to be based on information from the latter module. Both modules must be integrated with the agent in which we will evaluate the work with. Depending on the current state of the agent, a translation module may have to be created, to connect the main memories of the agent to the memories in Cognitive Trust Modelling Module. Additionally the Trustful Action Suggestion Module must also be added onto the planner component of the agent, so that the suggestions will be taken into account. The overall architecture dependencies can be checked in Figure 2.

5 Evaluation

The modules will be evaluated by integrating them with an agent capable of acting on a specific conflict or collaboration scenario, and then compare trust measurements according to the evaluation steps described below. While we just describe one scenario in this report we hope to be able to perform the evaluation steps in other scenarios as well.

5.1 Evaluation Steps

Evaluation will be performed through individual user testing of the agent, and we will try to gather at least 30 participants to ensure statistical viability. The testers will then be separated into two equally distributed groups, which we will designate by *Group A* and *Group B*. *Group A* will be the control group. The following steps will be performed for each user:

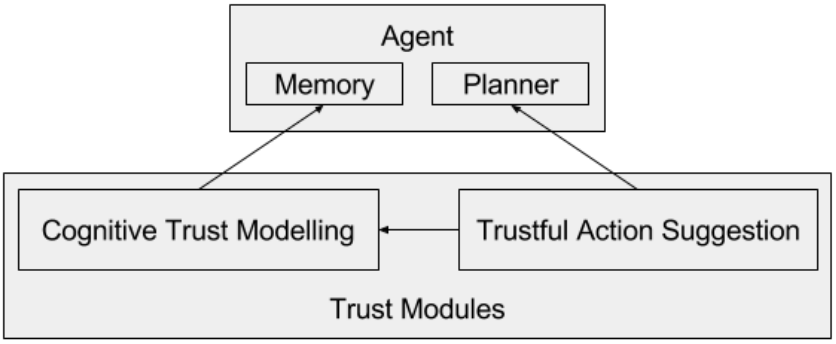

Figure 2: Solution architecture schematic

1. Perform a series of runs in a game-like scenario with an individual user and the agent as players; in *Group A* the agent will play **without** our modules and in *Group B* with them.
2. After interaction with the agent, the user will fill out a questionnaire to assert the value of Trust has in the Agent, in a range from 0 to 100; the questionnaire to be used will be the described in Section 3.2; whether we will use the complete version or the one described in is still to be decided.

We will then compare the averaged Trust value of both groups, and if the value of *Group B* is greater by a significant margin, it will provides positive feedback to the decision making module. Additionally we will check how closely did the questionnaire answers matched with the model created in the agent by the user trust module , providing a measure of accuracy of the model created.

5.2 *Split or Steal* scenario

As stated in Section 1, the project's evaluation will be done by integrating the developed modules in a agent, now finishing development, that is capable of acting as a player in the *Split or Steal* scenario, introduced in the British television show *Golden Balls* [19]. The scenario involves two players and stands as follows:

1. A large sum of money is prized for the game;
2. Each player receives two balls, one has 'Steal' written inside while the other has 'Split';
3. The balls can be stealthily open by the players, so only each player knows is written in what ball;
4. The players then have some time to discuss and negotiate between one another;

5. Finally the players choose one of their balls and show their content simultaneously and the game finalizes, giving one of the following results:
   - If both players choose 'Split', they split the prize money evenly;
   - If one picks 'Steal' and the other 'Split', the stealer gets all the prize money;
   - If both pick 'Steal', they both lose all the prize money.

From a game theory standpoint, this scenario is a variation on *Prisoner's Dilemma*, described in Section 2.3.1, with a payoff matrix shown in Table 2. But in this game, *Defecting* only weakly dominates *Cooperating*, because if an opposing player picks 'Steal' we get nothing whether we pick 'Steal' or 'Split', so 'Steal' only dominates if the opposing player picks 'Split' [37]. In the regular *Prisoner's Dilema* scenario, there's a sense of fear that pushes the players to *Defect*, as *Cooperating* will always present a worse personal result, regardless of the action chosen by the opponent. In this 'weaker' scenario, that is lessened, as 'Defecting' will not always wield a better result and so the player will be less likely to fear losing out by choosing the worse option.

	C_2	D_2
C_1	1,1	0,2
D_1	2,0	0,0

Table 2: Split or Steal Payoff Matrix

The most interesting step for evaluation is Step 4 of the scenario, the negotiation phase, where is most probable for the agent and user to perform some spoken interaction.

6 Planning

7 Conclusion

We started by introducing

Acknowledgment

The author would like to thank...

References

1. Simpson, J.a.: Foundations of interpersonal trust. In: Social psychology: Handbook of basic principles (2nd ed.). (2007) 587–607

2. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edition. (2009)
3. Grosz, B.J.: Collaborative Systems. AI Magazine (1996) 67–85
4. Allen, J., Ferguson, G.: Human-machine collaborative planning. International NASA Workshop on Planning (2002) 1–10
5. Allen, J., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Taysom, W.: PLOW : A Collaborative Task Learning Agent. Interpreting **22** (2007) 1514–1519
6. Bradshaw, J.M., Feltovich, P., Johnson, M.: Human-Agent Interaction. Handbook of HumanMachine Interaction (2011) 293–302
7. Lee, J.D., See, K.A., City, I.: Trust in Automation : Designing for Appropriate Reliance. **46**(1) (2004) 50–80
8. Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. (mar 1998)
9. Lashkari, Y., Metral, M., Maes, P.: Collaborative interface agents. AAAI '94: Proceedings of the twelfth national conference on Artificial intelligence **1** (1994) 444–449
10. Bickmore, T.W., Picard, R.W.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer-Human **12**(2) (2005) 293–327
11. Goodrich, M.a., Schultz, A.C.: Human-Robot Interaction: A Survey. Foundations and Trends® in Human-Computer Interaction **1**(3) (2007) 203–275
12. van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Haselager, W.F.G.: Do Robot Performance and Behavioral Style affect Human Trust ? International Journal of Social Robotics (2014)
13. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**(10) (oct 1992) 1243–70
14. Jones, S., Marsh, S.: Human-computer-human interaction. ACM SIGCHI Bulletin **29**(3) (jul 1997) 36–40
15. eBay Inc.
16. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: cognitive anatomy, social importance, and quantification. Proceedings of the International Conference on Multi Agent Systems (1998) 72–79
17. Rao, A.S., Georgeff, M.P.: BDI agents: From theory to practice. Icmas **95** (1995) 312–319
18. Sabater, J., Paolucci, M., Conte, R.: Repage: REPutation and ImAGE among limited autonomous partners. Jasss **9**(2) (2006) 117–134
19. Wikipedia: Golden Balls: https://en.wikipedia.org/wiki/Golden_Balls
20. Rousseau, D., Sitkin, S., Burt, R., Camerer, C.: Not so different after all: A cross-discipline view of trust. Academy of Management Review **23**(3) (1998) 393–404
21. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review **24**(1) (2005) 33–60
22. Castelfranchi, C., Falcone, R.: Trust Theory. 1 edn. John Wiley & Sons, Ltd, Chichester, UK (mar 2010)
23. Gambetta, D.: Can We Trust Trust? In: Trust: Making and Breaking Cooperative Relations. Blackwell (1988) 213–237
24. Marsh, S.P.: Formalising Trust as a Computational Concept. PhD thesis (apr 1994)
25. Abdul-rahman, A., Hailes, S.: Supporting Trust in Virtual Communities. System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on **00**(c) (2000) 1–9

26. Sabater, J., Sierra, C.: 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 (2002) 475
27. Huynh, T.D., Jennings, N.R., Shadbolt, N.R.: An integrated trust and reputation model for open multi-agent systems. Autonomous Agents and Multi-Agent Systems **13**(2) (2006) 119–154
28. Pinyol, I.: Reputation-Based Decisions for Cognitive Agents (Thesis Abstract). Doctoral Mentoring Program (Aamas) (2009) 33
29. Nash, J.: Non-Cooperative Games. The Annals of Mathematics **54**(2) (sep 1951) 286
30. Granatyr, J., Botelho, V., Lessing, O.R., Scalabrin, E.E., Barthès, J.P., Enembreck, F.: Trust and Reputation Models for Multiagent Systems. ACM Computing Surveys **48**(2) (oct 2015) 1–42
31. Han Yu, Zhiqi Shen, Leung, C., Chunyan Miao, Lesser, V.R.: A Survey of Multi-Agent Trust Management Systems. IEEE Access **1** (2013) 35–50
32. Pinyol, I., Sabater-Mir, J.: Computational trust and reputation models for open multi-agent systems: a review. Artificial Intelligence Review **40**(1) (jun 2013) 1–25
33. Noorian, Z., Ulieru, M.: The State of the Art in Trust and Reputation Systems: A Framework for Comparison. Journal of theoretical and applied electronic commerce research **5**(2) (aug 2010) 97–117
34. Huang, H., Zhu, G., Jin, S.: Revisiting Trust and Reputation in Multi-agent Systems. Computing, Communication, Control, and Management, 2008. CCCM '08. ISECS International Colloquium on **1** (2008) 424–429
35. Sutcliffe, A., Wang, D.: Computational Modelling of Trust and Social Relationships. Journal of Artificial Societies and Social Simulation **15**(1) (aug 2012) 523–531
36. Schaefer, K.: The Perception and Measurement of Human-Robot Trust. PhD thesis (2009)
37. Rapoport, A.: Experiments with N-Person Social Traps I: Prisoner's Dilemma, Weak Prisoner's Dilemma, Volunteer's Dilemma, and Largest Number. Journal of Conflict Resolution **32**(3) (sep 1988) 457–472

Appendices

A The Perception and Measurement of Human-Robot Trust: Items Table

Items
Act consistently
Protect people
Act as part of the team
Function successfully
Malfunction
Clearly communicate
Require frequent maintenance
Openly communicate
Have errors
Perform a task better than a novice human user
Know the difference between friend and foe
Provide Feedback
Possess adequate decision- making capability
Warn people of potential risks in the environment
Meet the needs of the mission
Provide appropriate information
Communicate with people
Work best with a team
Keep classified information secure
Perform exactly as instructed
Make sensible decisions
Work in close proximity with people
Tell the truth
Perform many functions at one time
Follow directions
Considered part of the team
Responsible
Supportive
Incompetent
Dependable
Friendly

Items
Reliable
Pleasant
Unresponsive
Autonomous
Predictable
Conscious
Lifelike
A good teammate
Led astray by unexpected changes in the environment

Table 3: The Perception and Measurement of Human-Robot Trust: Items Table