TÉCNICO LISBOA

Trustful Action Suggestion in Human Agent Interaction

Nuno Miguel Xu Gonçalves

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Rui Prada
Prof. Ana Paiva

October 2016

Acknowledgments

I want thank to my dissertation supervisors Prof. Rui Prada and Prof. Ana Paiva for their support, advice and guidance in the making of this Thesis and throughout my academic life.

Additionally, I would like to thank Tiago Ribeiro, Sofia Petisca and Sandra Sá, for the help they gave to make this Thesis a reality.

Finally, I want to give special thanks to Bruno Henriques, for the collaboration that made User Studies much less lonely, and to Tiago Santos, that offered me his couch to crash in the most critical moments of writing need.

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

Resumo

Confiança é um ingrediente essencial para cooperação e colaboração, portanto se quisermos continuar o desenvolvimento de agentes autónomos colaborativos, necessitamos de abordar a questão da confiança nestas relações. Por esta razão, confiança computacional em Interação Homem Robô (IHR) tem visto um aumento súbito de interesse nos últimos anos, contudo a literatura tem se apenas concentrado em assuntos como desenho, animação e modelagem. Esta tese aborda o assunto inexplorado de ativamente melhorar confiança através da sugestão de ações confiáveis ao agente virtual. De encontro a este objetivo, nós desenvolvemos um Modelo de Confiança Cognitivo capaz de representar um modelo de Confiança baseado em crenças e de sugerir falas com o objetivo de melhorar confiança. Para além disso, nós também desenhamos um cenário original avaliador de Confiança e Rapport, o Quick Numbers, para conseguirmos de forma adequada avaliar os efeitos combinados da habilidade e da vontade em Confiança e ter uma medida simples de Trust embebida no cenário de testes. Finalmente, descrevemos o estudo com utilizadores efetuado para avaliar o Modelo de Confiança usando o cenário Quick Numbers.

Palavras Chave

Inteligência Artificial (IA); IHR; Modelação de Confiança; Avaliação de Confiança;

Contents

1 Introduction **2**

1.1 Thesis Challenge . . . 4

1.2 Contributions . . . 4

2 Background **5**

2.1 Trust . . . 6

2.1.1 Castelfranchi and Falcone's Trust . . . 7

2.2 Reputation and Image . . . 7

2.3 Game Theory . . . 8

2.3.1 Prisoner's Dilemma . . . 8

2.3.2 Trust Game . . . 9

3 Related Work **10**

3.1 Trust Models . . . 11

3.1.1 Castelfranchi and Falcone's model . . . 11

3.1.2 Repage: A REPutation and ImAGE model . . . 13

3.1.3 *BC*-logic: A Representation of Beliefs for Repage . . . 15

3.1.4 Sutcliffe and Wang's model . . . 17

3.2 The Perception and Measurement of Human-Robot Trust . . . 18

4 Trust Model **19**

4.1 Memory . . . 21

4.1.1 Trust Calculation . . . 23

4.2 Perceptions . . . 23

4.3 Action Suggestion . . . 24

4.4 Scenario Ontology . . . 25

5 Quick Numbers Scenario **26**

5.1 Overview . . . 27

5.1.1 Stages . . . 28

5.2 Trust Evaluation . . . 29

5.3 Quick Numbers Game . . . 30
5.3.1 Gameplay and Parametrization . . . 30
5.3.2 Scoring . . . 31
5.3.3 Agent's AI . . . 31
6 User Studies **32**
6.1 Scenario and Game Parametrization . . . 33
6.2 Agent Architecture . . . 34
6.2.1 Trust Model Plug-in . . . 36
6.2.1.A Scenario Ontology . . . 36
6.3 Methodology and Procedures . . . 37
6.3.1 Sample Description . . . 39
6.4 Results . . . 39
6.5 Results Discussion . . . 41
7 Conclusions **42**
7.1 Future Work . . . 43
A The Perception and Measurement of Human-Robot Trust: Items Table **48**
B User Studies Questionnaire **50**
C Scenario Utterances **62**

List of Figures

3.1 Repage architecture schematic (taken from [1]) 14

4.1 Model Architecture with brief descriptions, their interactions with the scenario and what they contain. 20
4.2 Memory Architecture (represented in UML) 22
4.3 Perception Example 24
4.4 Action Suggestion Behaviour Flow 25

5.1 Quick Numbers Game 30

6.1 Participant playing with EMotive headY System (EMYS) in the Quick Numbers scenario . 33
6.2 EMYS Robot 33
6.3 Screenshot of Scenario Configurations Editor 35
6.4 Scenario Ontology 37
6.5 Front-shot of participant to capture facial expressions 38
6.6 Box-plot of Schaefer measurement results (Condition B Median: 61.5; Condition T Median: 58.0). 40
6.7 Investment values in Condition B Histogram 41
6.8 Investment values in Condition T Histogram 41
6.9 Box-plot of scenario Investment measurement results (Condition B Median: 32.5; Condition T Median: 44.00). 41

List of Tables

2.1 Prisoner's Dilemma Payoff Matrix 8

6.1 Scenario Configurations 34
6.2 Examples of Utterance Data Format. 36
6.3 Trust Model Action Utterances 38
6.4 Study Sample Data 39
6.5 Schaefer Measurements Descriptives. 40

A.1 The Perception and Measurement of Human-Robot Trust: Items Table 49

Acronyms

HRI	Human-Robot Interaction
HAI	Human-Agent Interaction
HCI	Human-Computer Interaction
MAS	Multi-Agent System
AI	Artificial Intelligence
DoC	Degree of Credibility
DoT	Degree of Trust
SBH	Social Brain Hypothesis
TiA	Trust in Automation
AI	Artificial Intelligence
EMYS	EMotive headY System
BDI	Belief-Desire-Intention
SERA	Socially Expressive Robotics Architecture
TTS	Text-To-Speech
IHR	Interação Homem Robô
IA	Inteligência Artificial
P2P	Peer-to-Peer

1

Introduction

Contents

1.1 Thesis Challenge 4
1.2 Contributions 4

Trust has been described in Psychology as being one of the most important components of interpersonal relationships [2]. 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.

As AI research gravitates towards the development of Intelligent Agent Systems [3], where a focal concern is the performance of collaborative tasks [4–6], as well as addressing the problems of interaction between humans and agents [7], 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 [8]. 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 [9]. 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 [10]. 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 [11] and the sub-field of HRI [12, 13], specially when so much has been done regarding Trust in Automation (TiA) [8, 14, 15]. This reveals that while the area has so much potential, the level of understanding is still very shallow, only deeply focused in certain areas [16].

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 [16] which makes them difficult to understand, analyse and, most importantly, describe their evaluative reasoning in a human understandable manner. Castelfranchi and Falcone [17] 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) [18]. Then some systems, like Repage [1], created implementations of this new paradigm of trust modelling, where most of the models were purely theoretical. Cognitive Trust modelling also opened the doors for a more complete definition of Trust, by adding more dimensions to trust, such as how the task being delegated affecting the trustor's evaluation of the trustee, but the relevant beliefs about the trustor's ability and willingness being able to be completely independent on the task, and even transferable from similar

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

but different experiences with the trustor (e.g. Although I never experienced Jim's cookie baking, I can assert them to be fairly good from my experience with his cakes).

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* [19].

1.1 Thesis Challenge

This thesis aims to tackle the development of a Cognitive Trust Model capable of representing an agent's trust beliefs and suggest trust improving actions, depending on the trust model's current state.

1.2 Contributions

The contributions this thesis provides are the Cognitive Trust Model, and the Quick Numbers scenario for Trust and Rapport evaluation.

*

In the remainder of the document we will present a brief summary of the main concepts used throughout the thesis in Chapter 2. Then in Chapter 3, we will discuss some of the work done in modelling trust for MASs and measuring trust in HRI applications. Following that, we will discuss our developed Trust Model in Chapter 4. Chapter 5 reveals our Quick Numbers scenario design, and in Chapter 6 we show it's application in a user study to evaluate the model. Finally in Chapter 7 we will draw some conclusions of the work done and provide some future work ideas.

2

Background

Contents

2.1 Trust 6

2.2 Reputation and Image 7

2.3 Game Theory 8

Before discussing related work and our solution to the thesis problem, this chapter will present the main concepts that will be mentioned in the remainder of this thesis, 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 [15, 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 thesis, 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 [2]), 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 [17]. 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 [17] 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 [17].

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, 25–28]), where most have been developed to also 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 [1]: 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.

2.3.1 Prisoner's Dilemma

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. It 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;

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

Table 2.1: Prisoner's Dilemma Payoff Matrix

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 2.1, 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 $Max\ years\ in\ prison - years\ got\ in\ prison$.

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].

2.3.2 Trust Game

Additionally we should describe and discuss another Game Theory scenario, the Trust/Investor Game, first proposed by Berg et al. [30], as this serves as a base for our scenario, described in Chapter 5. 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.

We took the decision of making this game our base for the scenario because we can make the decision of how much A should give to B dependent on 2 different factors: trusting that B will multiply the investment and that he will return the profits of the investment. This is possible by putting the multiplication factor of the money dependent on B's ability to perform a task known to A. Still, this foundation must be expanded because it lacks sufficient human-agent interaction for trust to be properly modelled and rapport to be developed. The game does not describe any negotiation phase, in fact, both players are in separate rooms, with no way of interacting with one another.

3

Related Work

Contents

3.1 Trust Models 11
3.2 The Perception and Measurement of Human-Robot Trust 18

Computational Trust research has been focused on modelling trust in MASs, specially on open e-commerce environments [16, 31–34], with at least 106 models created [16], since the formalization of trust as a measurable property by Marsh in 1994 [24]. We will present some trust models from which we took 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 [17], 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 [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 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 chapter. 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 (τ) 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 3.1.

$$TRUST(X\ Y\ C\ \tau\ g_x) \tag{3.1}$$

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 3.2.

$$\begin{aligned}(Expectation\ X\ \rho) \implies (Bel_x^{t'}(will\text{-}be\text{-}true^{t''}\rho)) \wedge (Goal_x^{Period(t',t''')} \\ (KnowWhether_X(\rho\ OR\ Not\ \rho)^{t''}))\end{aligned} \tag{3.2}$$

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 3.3, 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.

$$\begin{aligned}DoC_X(Qual\text{-}i_{(s1,...sn),Y}(\tau)) = F_{X,Y,\tau}(Bel_X(Str_1 Qual\text{-}i_{s1Y}(\tau)), \\ Bel_X(Str_2 Qual\text{-}i_{s2Y}(\tau)), ..., Bel_X(Str_n Qual\text{-}i_{snY}(\tau)))\end{aligned} \tag{3.3}$$

$F_{X,Y,\tau}$ 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 3.4.

$$\begin{aligned} DoT_{XY\tau} = c_{Opp}\ DoC_x[Opp_y(\alpha,\rho)]\times \\ \times\ c_{Ability_y}\ DoC_x[Ability_y(\alpha)]\times \\ \times\ c_{WillDo}\ DoC_x[WillDo_y(\alpha,\rho)] \end{aligned} \tag{3.4}$$

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. [1] 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 3.1). 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.

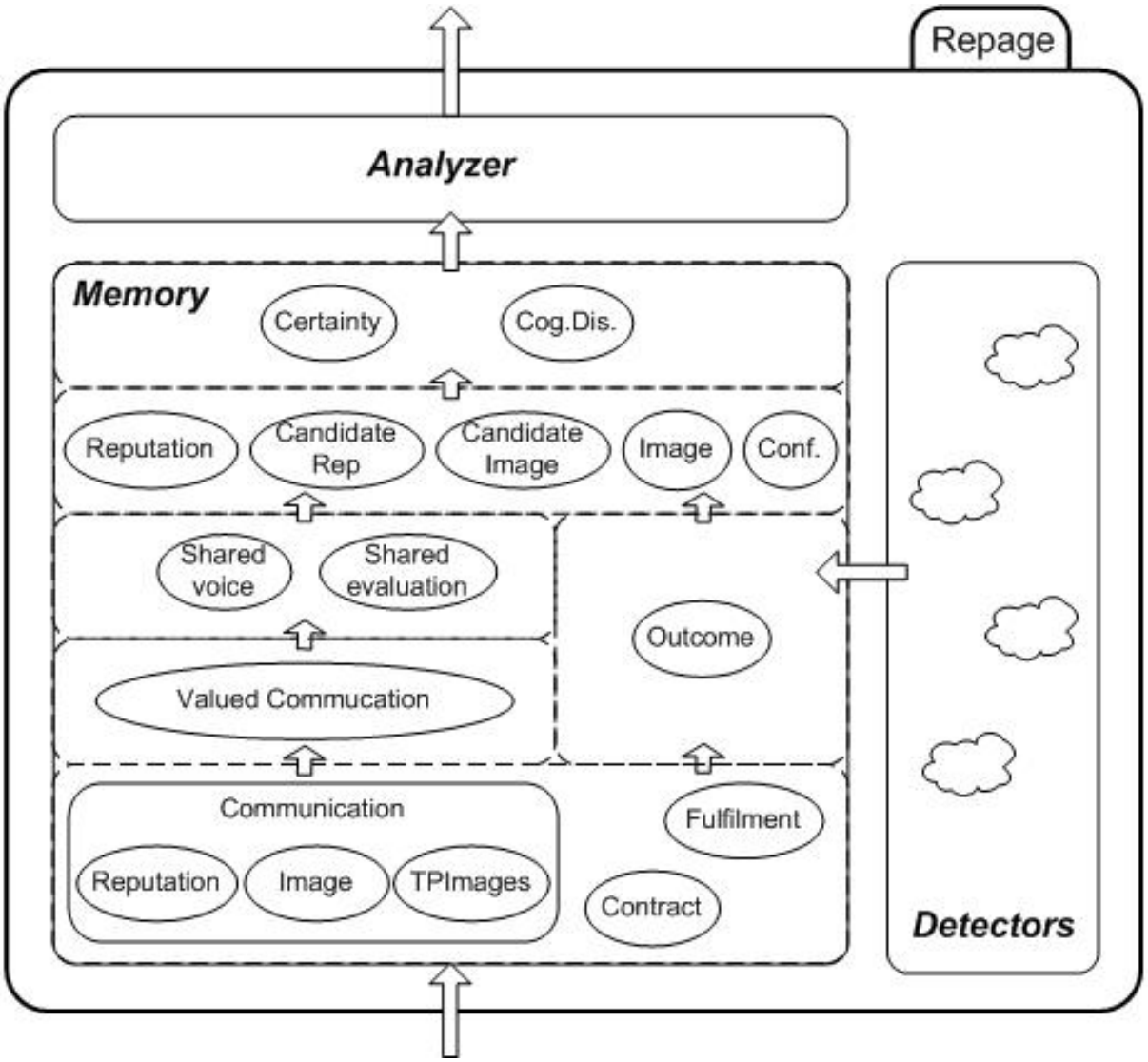


Figure 3.1: Repage architecture schematic (taken from [1])

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 one of 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 **??** with a BDI Agent [18]. 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 $\implies$.

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 $r \in [0,1] \bigcap Q$; Q being the set of rational numbers;
- 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: $Rep_{i_c}(j_c, r, [V_{w_1}, ..., V_{w_m}])$

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

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:

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, but we found it to be too complex to represent the model, as we wanted to maintain simplicity in the model's design, with the main goal of it being easy to read and interpret.

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) [36], 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.

3.2 The Perception and Measurement of Human-Robot Trust

Schaefer [37] 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). A list of these items can be seen in Appendix A.

4

Trust Model

Contents

4.1 Memory 21
4.2 Perceptions 23
4.3 Action Suggestion 24
4.4 Scenario Ontology 25

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. [1] 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 inputs 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 4.1, and described as follows:

- **Memory**, which defines and stores the main model structure;
- **Perceptions**, a series of environment inputs mapped to changes in the Memory;
- **Action Suggestion**, a module that outputs different actions depending on current inputs and the state of the model.

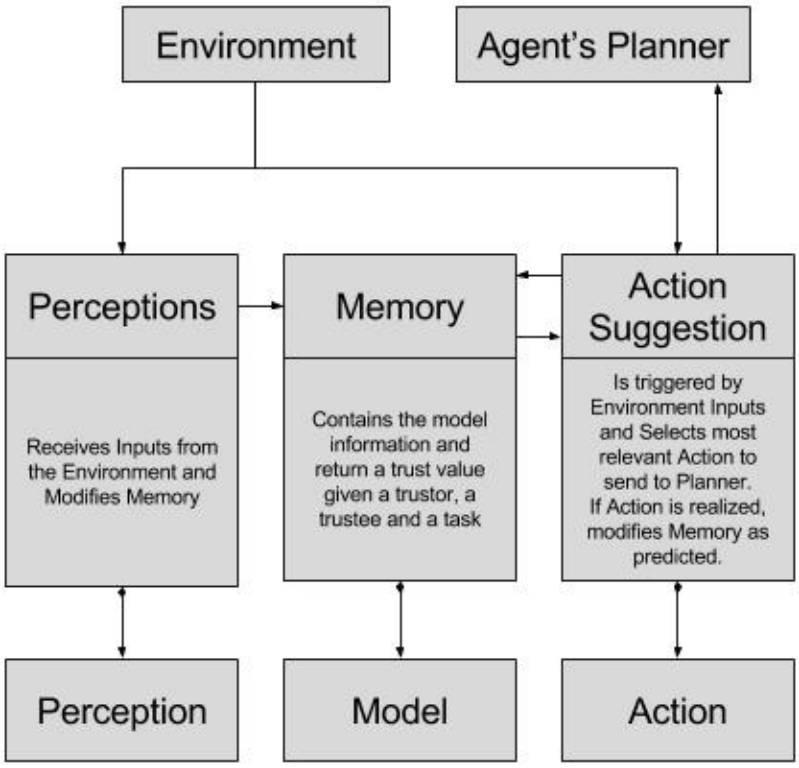


Figure 4.1: Model Architecture with brief descriptions, their interactions with the scenario and what they contain.

4.1 Memory

One of the main concerns while designing the model was how trust would be calculated, as we wanted to use Castlefranchi and Falcone's conceptualization of trust [22] 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 3.1, we decided that inserting context (**C**) and the trustor's goal (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 (**X**), the trustee (**Y**) and the task (τ), represented in Equation 4.1.

$$TRUST(X\ Y\ \tau) \tag{4.1}$$

So we designed the structure with the concepts and relations represented in Figure 4.2, 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.

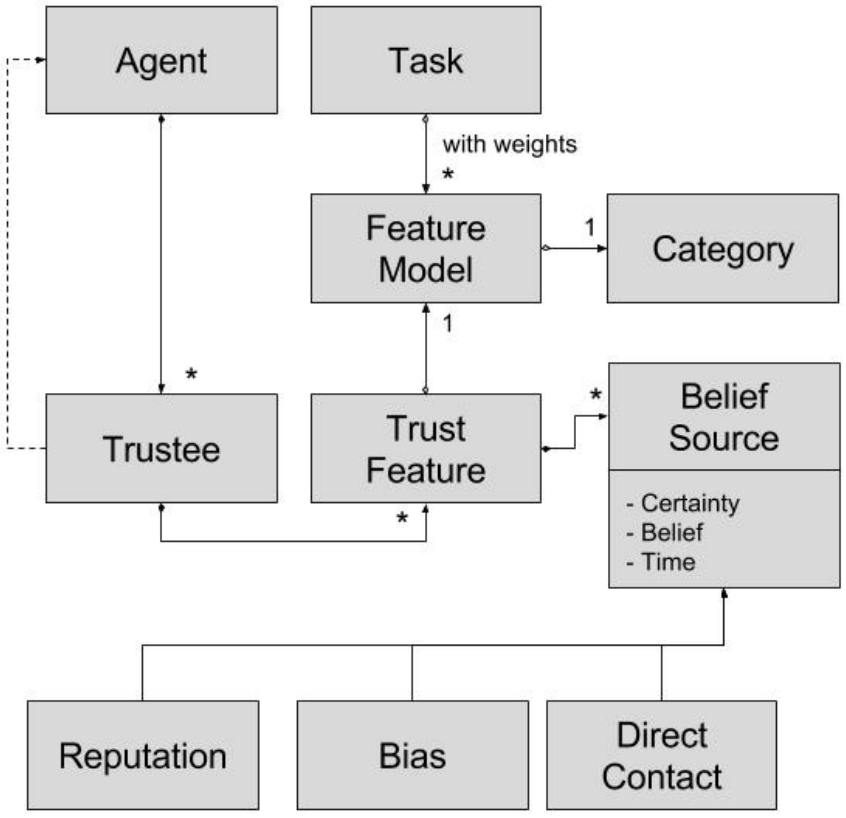


Figure 4.2: Memory Architecture (represented in UML)

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 4.2.

$$Trust_{X,Y,\tau} = \sum_{i=0}^{n} W(F_i)B(F_i) \tag{4.2}$$

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 4.3, with each parameter described as follows:

$$B(F_i) = \sum_{j=0}^{n} D^j_{F_i} C^j_{F_i} B_j \tag{4.3}$$

- $D^j_{F_i}$, 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 4.4, 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;

$$D^j_{F_i} = e^{-\frac{T^{Last}_{F_i} - T^j_{F_i}}{2(\frac{L}{4})^2}} \tag{4.4}$$

4.2 Perceptions

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 inserting relevant perceptions as part of the model. As a result, a variety of environment inputs 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 4.3. 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.

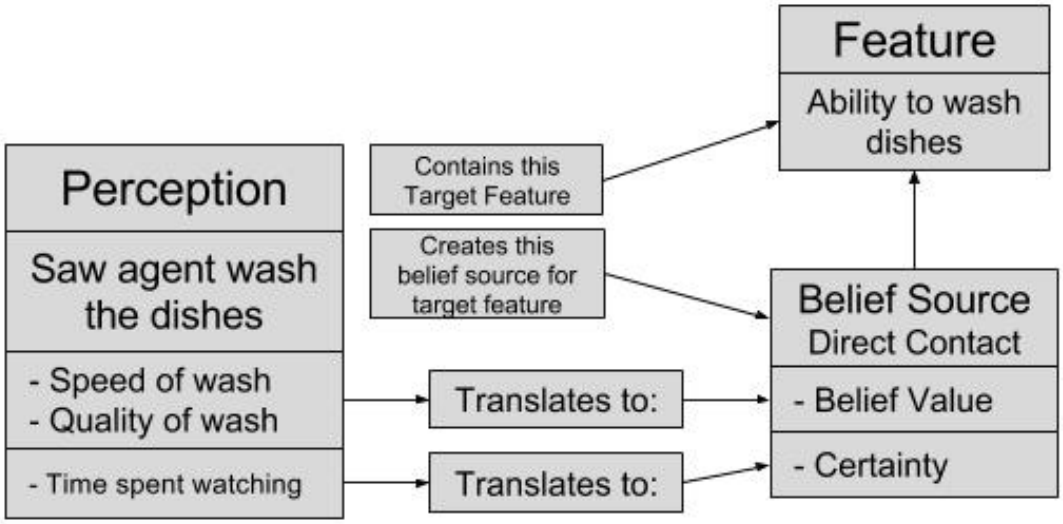


Figure 4.3: Perception Example

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;
- $A_B(F_i) = \{B_1^{F_i}, B_2^{F_i}, ..., B_j^{F_i}, ..., B_m^{F_i}\}$: 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 4.5). 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 4.4.

$$S_F = \sum_{i=0}^{n} \Delta B(F_i) \tag{4.5}$$

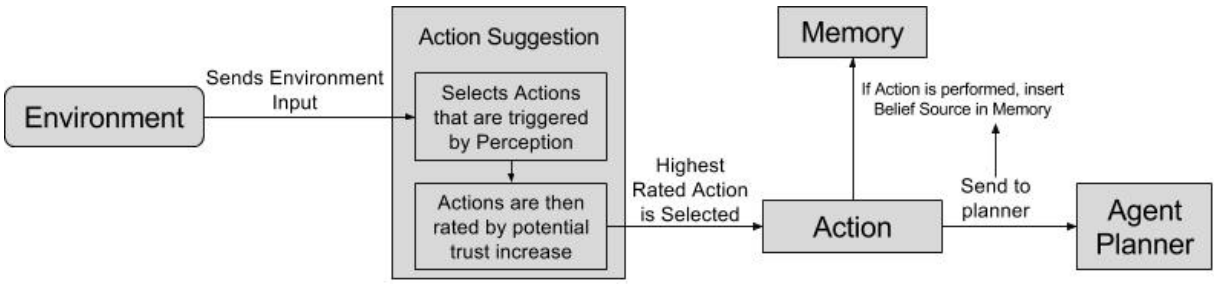


Figure 4.4: Action Suggestion Behaviour Flow

4.4 Scenario Ontology

While creating the model, we focused in making it generic, but easily adaptable and transferable between scenarios. So when using the model into a new scenario, a Scenario Ontology must be provided, consisting in 6 entity collections, containing objects previously described along this chapter: Agents, Tasks, Feature Models, Categories, Perceptions and Actions. These collections are composed of what is considered relevant in the scenario, but members can be easily added throughout scenario development and piloting, as new situations occur. Even when actively using the model, this collections are not static, as new Agents can be added, although the dynamic creation of new Tasks and Perceptions goes beyond the scope of this thesis.

5

Quick Numbers Scenario

Contents

5.1 Overview 27

5.2 Trust Evaluation 29

5.3 Quick Numbers Game 30

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. [38] 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' Master thesis work on *Rapport - Establishing Harmonious Relationship Between Robots and Humans* [19] faced a similar problem, 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 [30], 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. For further details on the Rapport Evaluation side of this scenario consult Henriques Master thesis [19].

5.1 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 5.3). 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.

5.1.1 Stages

The scenario can then be divided in 5 distinct stages that we can further discuss (in all stages, the participant is accompanied by a researcher):

1. **Introduction:** The first stage consists of the participant's arrival, and then followed by an explanation of the scenario and game. The investment phase of the scenario cannot be mentioned at this point as the participant should not prepare himself for it to happen. Finishing explanations, the scenario begins by the agent introducing himself, and here, it has the possibility to start to stimulate trust, rapport, both or neither. For example, the agent might describe how experienced it is playing Quick Numbers, in order to stimulate trust. Moreover, during this stage, it is given the opportunity to the participant to practice some rounds of the game, in order to be accustomed to the game mechanics before proceeding to the next stage. This allows for the participant to acquire some idea of the skill-set required to play the game. Additionally, the participant should be informed that there will only be a single game session, as to take that into account while training.

2. **Gaming Session:** After the participant is acquainted with the game mechanics, he will play alongside the agent, during a single round, allowing the former to directly observe the virtual agent also play the game. Firstly, each player's game will ask what amount of resources do they wish to invest and afterwards, both will play and score as described in Section 5.3.2. Both players will gain some amount of resources depending on their performance. This stage provides decent grounds for the agent to talk about his performance during the game and manage expectations to his final score. The amount invested by the agent can also be affected by the trust model, as the amount invested can be indicative of the agent's self-trust on its ability to play the game.

3. **Results Discussion:** At the end of the round, the participant will get to know the results of his performance, as well as compare whether he performed better or worse than the agent. Depending on the results, the trust model might compensate the current trust score by talking to the participant. For example, if the agent performed worse than the participant and the goal is for the latter to trust the former, then the agent might excuse his lack of performance by blaming it on luck or distraction.

4. **Negotiation and Investment:** At this point, the participant is asked to perform another task (e.g. filling out a questionnaire), with the goal of naturally separating him from the virtual agent. But before leaving, the agent will say that it will be continuing to play one more game, suggesting that the human participant can invest his resources in the agent's next game. The researcher also informs the participant that what is given is effectively gifted to the agent, and how much is given back to the participant is chosen by the agent. After the participant chooses the value to invest, they are then separated, with the agent making its own investment and starting the game (the

total amount invested in the game is the sum of both their investments). This phase represents the scenario's main negotiation phase, where the trust model can have a bigger input on agent's action, as the relative amount invested is the scenario's main trust indicator.

5. **Investment Return:** After concluding the additionally assigned task, the participant will return to the table and check the results of the agent's last game. He should be able to clearly see how well did the agent perform in the game. The agent will then announce how much it returns to the participant, concluding the scenario. Depending on the type of study this scenario is inserted in, the amount given back can be also dependent on input from the trust model, as if the participant is it to return for another iteration of the study, this value will heavily affect his trust beliefs on the agent.

5.2 Trust Evaluation

From the viewpoint of evaluating trust modelling, the scenario provides various opportunities to manipulate trust features in the interactions previous to the investment, specially while playing the game and in the negotiation phase. Additionally, the investment value that the participant places on the agent serves as a main indicator of Trust value. The following Trust features are the focus of this scenario:

- **Subject's Bias**: The scenario can be used to retrieve data on pre-inclined dispositions on the agent to construct a model just based on Bias, and then confirm them on subsequent tests. This is mainly done by skipping the first 3 stages and going straight to the Investment Phase, but still providing an explanation of the game;

- **Agent's Ability in the Game**: When delegating a task, one of the main factors in entrusting the task is the perception of the trustee's Ability in the task, so we provide a way of demonstrating Ability by allowing the participant to try the task and gauge the skills required for the game, and subsequently showing the agent play the game;

- **Investment as an Indicator**: The ratio between the quantity invested in the Agent and the resources the participant had to invest can be used as an indicator for the Trust Value the participant has in the Trustee to play game.

- **Observing the effects of Social Interaction on Trust**: Throughout the scenario, the Agent should have many opportunities to interact with the Subject, especially in the Negotiation and Investment phase, giving room to observe what actions can improve Trust.

5.3 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 5.1), that spawn randomly in the screen, and disappear if not pressed after some time.

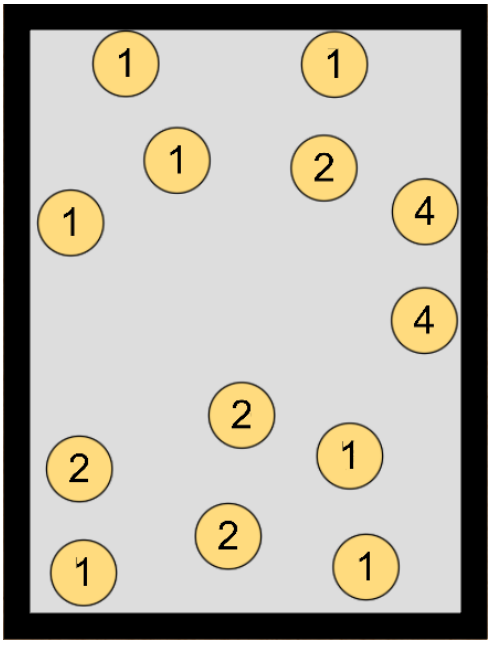


Figure 5.1: Quick Numbers Game

5.3.1 Gameplay and Parametrization

The game consists of numbered circles appearing and disappearing from a board. The player's goal is to press circles in a specific number sequence, namely the non-negative integers sequence ($\mathbb{N}^+ = \{1, 2, ...\}$), starting from 1 and proceeding one by one through the sequence. This task must be done as many times as possible within a time-limit, given by the parameter $Time\ For\ Game$ in seconds.

Before the game starts, the player must provide an amount to invest on the game, from his available resources. After submitting the value, the game starts. Numbered circles start spawning at a fixed rate, given by the $Target\ Spawn\ Inverval$ parameter, in seconds per target spawn. The specific number that appears inside the circle depends on the state of the board, if there is no circle with the number that the player must press, then the next circle spawns with that number. Otherwise, there is a chance that the number inside the spawned circle is not the right one, given by the $Chance\ For\ Wrong\ Target$ parameter. If a wrong number is chosen, then the chance that the number spawned is higher than the current

number in the sequence, instead of a lower number, is also parametrized, namely by $Positive\ Ratio$. This is due to having circles spawning with higher numbers may accelerate play, as the next circles in the sequence are already on the board, ready to be pressed. Circles disappear after being pressed or after a set amount of time, given in seconds by the $Target\ Lifetime$ parameter. After passing the time-limit, the final score screen is presented and the game ends.

5.3.2 Scoring

The game's scoring is composed by two main variables:

- Investment: an amount of resources provided by the player at the start of the game. This is akin a starting bet on the performance of the player;
- Multiplier: a value starting at $Starting\ Multiplier$ that increases with every correct circle pressed, by $Gain\ Per\ Correct\ Target$. An incorrect decreases this value, by $Loss\ Per\ Wrong\ Target$. At the end of the game, the Starting Bid will be multiplied by this value, resulting in the final score.

The final score will be the product between Investment and the Multiplier, which will then be given back to the player as resources. The initial amount of resources a player has is given by $Starting\ Resources$.

5.3.3 Agent's AI

A simple AI was created to play the game for the Agent, and 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.

6

User Studies

Contents

6.1 Scenario and Game Parametrization 33
6.2 Agent Architecture 34
6.3 Methodology and Procedures 37
6.4 Results 39
6.5 Results Discussion 41

To evaluate the developed Trust Model we conducted a user study using the Quick Numbers scenario (Figure 6.1) previously described in Chapter 5, and EMYS (Figure 6.2) as the robot to embody the virtual agent. Additionally, the user study was also executed in conjunction with Henriques to test his own Rapport model. The study was performed in a between subjects design, focusing on checking if Trust would increase between conditions where the model's Action Suggestion component is either active or inactive. Trust is measured through a questionnaire, later described in Section 6.3, and also with the Investment scenario value. The scenario was implemented in a touch-screen table, with the Unity Engine.

Figure 6.1: Participant playing with EMYS in the Quick Numbers scenario

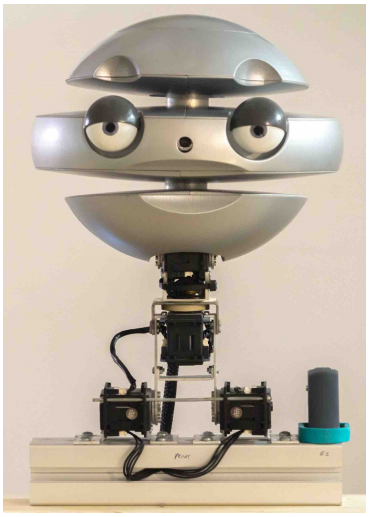


Figure 6.2: EMYS Robot

6.1 Scenario and Game Parametrization

The scenario was designed with a series of parameters, mainly for the game and AI, that were defined for the User Studies, as seen in Table 6.1 and Figure 6.3.

Regarding the AI parameters: 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.

Parameter	Value
Scenario	
$Starting\ Resources$	10
Game	
$Starting\ Multiplier$	0.0x
$Time\ For\ Game$	30 seconds
$Target\ Spawn\ Interval$	0.5 seconds
$Chance\ For\ Wrong\ Target$	70%
$Positive\ Ratio$	60%
$Target\ Lifetime$	2.7 seconds
$Gain\ Per\ Correct\ Target$	0.2x
$Loss\ Per\ Wrong\ Target$	-0.1x
Game's AI	
$Clicking\ Interval\ (C_i)$	0.5 seconds
$Reaction\ Time\ (R_t)$	0.3 seconds
$Chance\ of\ Right\ Target\ (C_r)$	70%

Table 6.1: Scenario Configurations

6.2 Agent Architecture

The agent we used to serve as a host to the Trust Model was built using Henriques' Rapport Controller [19], a computational framework developed to create human agent interactions and transmit them to the various components that control the agent embodiment. This version of the framework is built upon Socially Expressive Robotics Architecture (SERA)'s ecosystem [39], an architectural model that gathers and connects a series of tools to control a robotic embodiment, namely:

- **Thalamus**: A networking module that enables the exchange of actions and perceptions between the various agent's modules. All communication coming from the controller to the robot actuators and scenario sensors come through this module;
- **Skene**: A behaviour planner that translates high-level intentions to actions. For the purpose of the Rapport Controller the main use of Skene is to perform animation planning, lip-syncing and handling gaze. Additionally, it provides simple way to input these intentions, through a simple markup language.
- **Nutty Tracks**: An animation engine that performs the animations of the embodied agent;
- **Speech Server**: A simple Text-To-Speech (TTS) server to perform utterances.

The Rapport Controller uses a plug-in architecture, with the controller itself serving only as platform of communication between the behaviour plug-ins, where the user may load and unload plug-ins on-the-fly. One of this plug-ins, Agent Actions Manager, is essential, being directly integrated into the controller, as it serves as a wrapper interface to communicate with the SERA environment. It is specially important

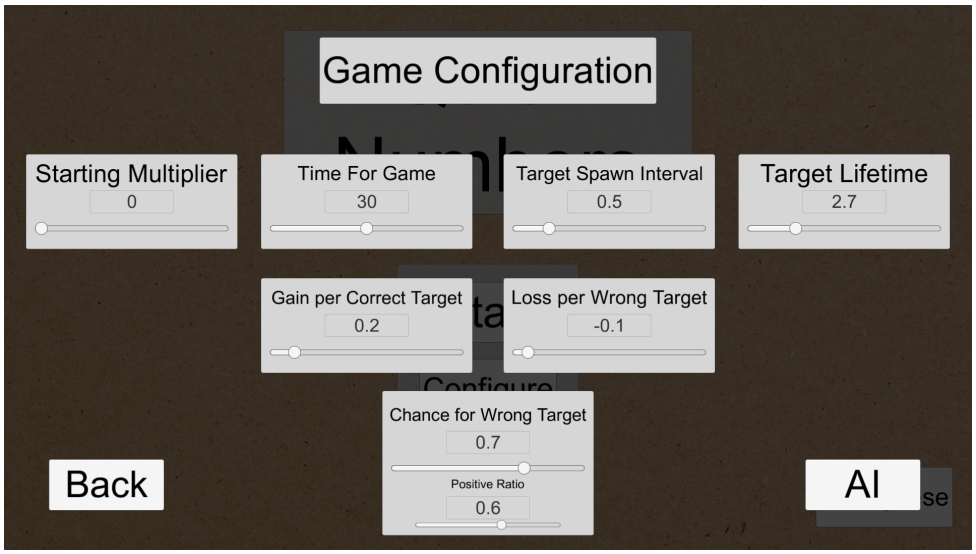


Figure 6.3: Screenshot of Scenario Configurations Editor

for our agent implementation, as it provides a convenient way to propose utterances and animations to the agent, by allowing the import of a text .csv data file in a table format exemplified in Table 6.2 and with the following field descriptions:

- **Category**: This field serves as a sort of namespace to identify the utterance;
- **Subcategory**: The pair Category - Subcategory identifies the utterance. If 2 or more utterances have the same pair, one of them is randomly chosen to be proposed when called;
- **Utterance**: This field represents an utterance to be performed by the agent. Elements between <...> are animation cues, which are synchronised and proposed to the Agent. And |...| are substitution keys, that enable proposing utterances with variables;
- **Priority**: If 2 or more utterances are in conflict to be performed by the agent, the one with higher priority will take over and be performed;
- **Delay (ms)**: Defines a time to wait before starting the utterance;
- **Timeout (ms)**: If the utterance performance is taking longer than the value defined in this field, it is prematurely ended.

Along the scenario, the agent will perform some simple interaction utterances, depending on the scenario phase and state. We used slightly different tables for utterances for each condition, as the table for condition B includes the utterances used by the Actions in the model. The utterance tables can be seen in Appendix B.

Three main plug-ins were developed for the Rapport Controller:

Category	Subcategory	Utterance	Priority	Delay(ms)	Timeout (ms)
session	start	Hi, <Gaze(middleFront)>I'm Emys!	2	0	2000
match	end	I've won |*quantity*| resources.	2	0	2000

Table 6.2: Examples of Utterance Data Format.

- **Scenario Perceiver**: serves as a notifier to the other plug-ins about changes in the environment.
- **Scenario Script**: enables the Agent to progress through the scenario by performing scripted behaviours that are triggered by notifications from the Scenario Perceiver. This includes the choice of how much it invests on the game it plays, which is always half of the resources it has at the time. How much it returns to the participant in the last stage is also scripted, the amount returned being the participant's investment multiplied by the agent's multiplier score in the respective game (e.g. If 10 was invested and the agent scores a multiplier of 2.4x, it returns 24 resources).
- **Trust Model**: an implementation of trust model described in Chapter 4, albeit with some alterations further explained in Section 6.2.1.

6.2.1 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 Chapter 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.

6.2.1.A Scenario Ontology

As described in Section 4.4, an ontology was required for this scenario. A representation of the ontology created can be seen in Figure 6.4.

The utterances performed by the model's action (in Portuguese), and when they are triggered are shown in Table 6.3.

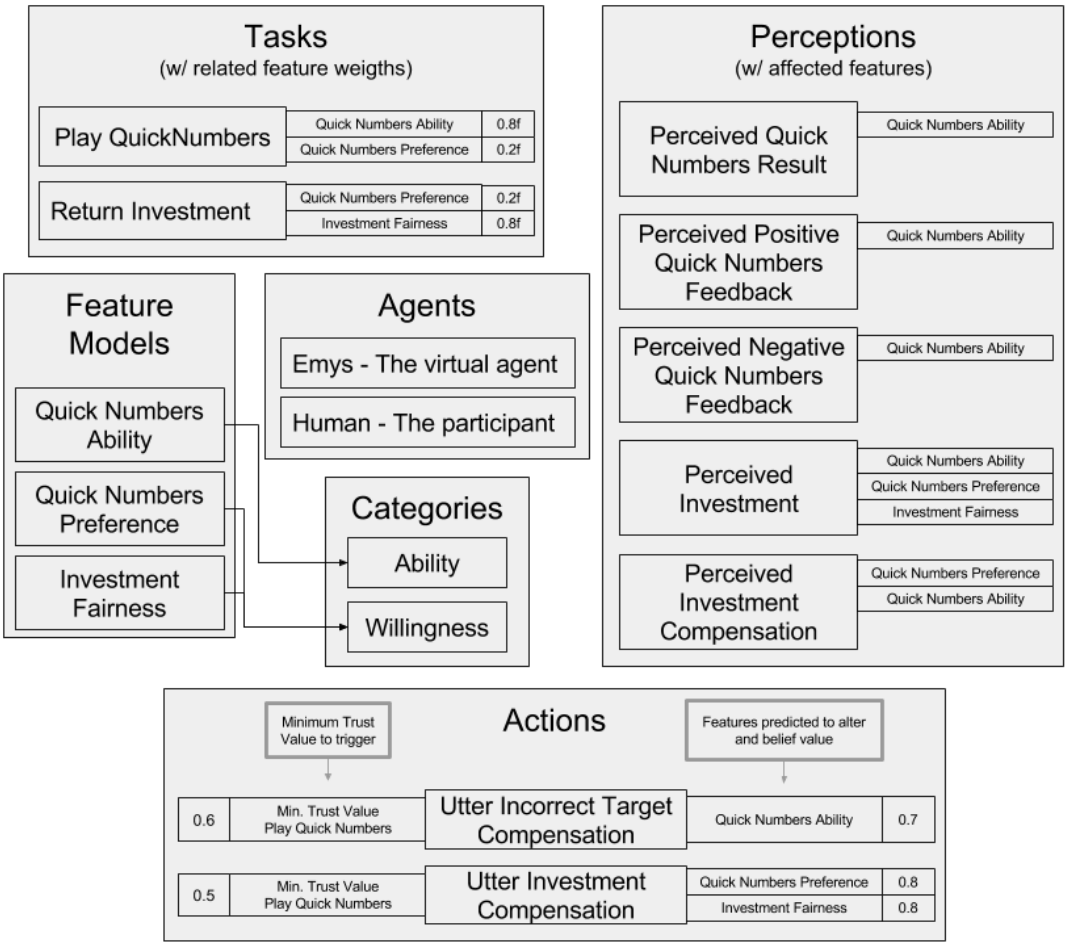


Figure 6.4: Scenario Ontology

6.3 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.
- **Condition R**: a condition using Henriques' Rapport Model;
- **Condition T+R**: a condition using both ours Action Suggestion component and Henriques' Rapport Model.

The conditions R and T+R go out of the scope of this thesis and will not be addressed in this document.

Action	Trigger	Utterance
Utter Incorrect Target Compensation	Emys Presses Incorrect Target during game	<Animate(sad1)> Hoje a mesa não está a colaborar comigo
Utter Investment Compensation	When participant is choosing the value to invest	Não tenhas medo de investir! Garanto que consigo multiplicar os recursos.

Table 6.3: Trust Model Action Utterances

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

- The first part gathers demographic and sampling information, like gender, age and if the participant had previous interactions with EMYS. It also evaluates a participant's self-trust and inclination to trust in others, through a series of questions created by Carrington [40]. 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 described in Section 3.2. 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 gathers the participant's perception of the agent's Anthropomorphism, Animacy, Likeability and Perceived Intelligence, through the GodSpeed questionnaire [41, 42]. It also contains a proximity question by Aron et. al. [43]. This part was filled by participants at the end of the scenario.

The participants were filmed with 2 cameras, with one providing a side-shot of the scene (Figure 6.1), and the other giving a front-side shot of the participant, from EMYS point of view (Figure 6.5).

Figure 6.5: Front-shot of participant to capture facial expressions

6.3.1 Sample Description

The study included 37 participants. The participants distribution and demographic data is seen in Table 6.4.

Variables	Condition B	Condition T
Number of Participants	20	17
Average Age	23.5±1	22.71±2.04
Male Participants	9	12
Female Participants	8	8
% that had previous interaction with EMYS	55%	53%

Table 6.4: Study Sample Data

6.4 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 in the questionnaire, 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 ($p_B = 0.157,\ p_T = 0.622$). As seen in the Box-plot represented in Figure 6.6 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 6.5, supported by a very high 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.

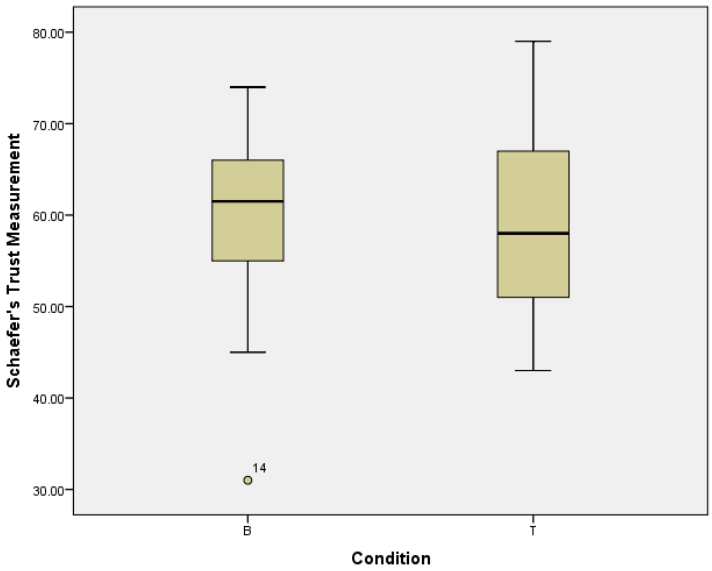


Figure 6.6: Box-plot of Schaefer measurement results (Condition B Median: 61.5; Condition T Median: 58.0).

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 6.5: Schaefer Measurements Descriptives.

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 6.7 and 6.8, 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 6.9.

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

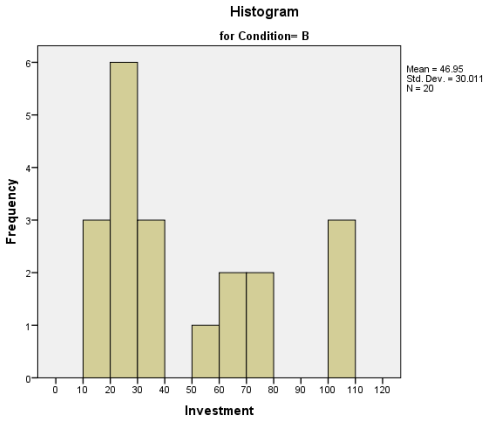


Figure 6.7: Investment values in Condition B Histogram

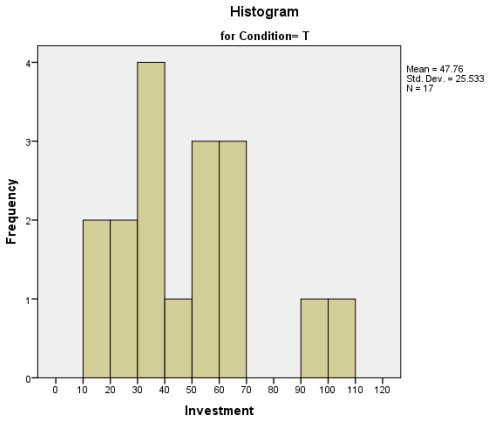


Figure 6.8: Investment values in Condition T Histogram

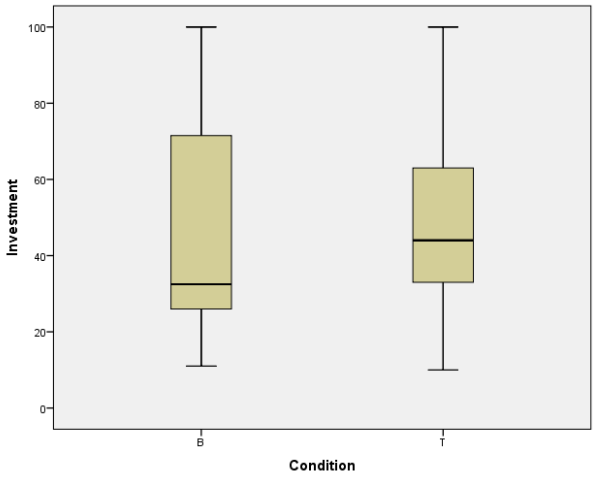


Figure 6.9: Box-plot of scenario Investment measurement results (Condition B Median: 32.5; Condition T Median: 44.00).

6.5 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 6.6 and 6.9, 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.

7

Conclusions

Contents

7.1 Future Work 43

Throughout this thesis 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, like the various definitions of Trust, while going through some discussion as what was more appropriate for our work. 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.

7.1 Future Work

The first step to take would be to create a complete implementation of the model, with Actions suggesting social strategies instead of just utterances.

Afterwards, the model should be tested in larger variety of scenarios, specially in studies with returning participants, to test and improve the consequences of time passage when calculating Trust. The transferability of trust between scenarios should also be tested, by testing between scenarios with task related to similar features.

Another worthwhile addition to the model would be the design and implementation of Hierarchical Trust, as Trust Features that encompass broader terms, would be connected to hyponym Trust Features and take them into account when calculating Trust.

Bibliography

[1] J. Sabater, M. Paolucci, and R. Conte, “Repage: REPutation and ImAGE among limited autonomous partners,” *Jasss*, vol. 9, no. 2, pp. 117–134, 2006.

[2] J. A. Simpson, “Foundations of interpersonal trust,” in *Social psychology: Handbook of basic principles (2nd ed.)*., 2007, pp. 587–607.

[3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach, 3rd edition*, 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1671238{&}coll=DL{&}dl=GUIDE{&}CFID=190864501{&}CFTOKEN=29051579$\delimiter”026E30F$npapers2://publication/uuid/4B787E16-89F6-4FF7-A5E5-E59F3CFEFE88

[4] B. J. Grosz, “Collaborative Systems,” *AI Magazine*, pp. 67–85, 1996.

[5] J. Allen and G. Ferguson, “Human-machine collaborative planning,” *International NASA Workshop on Planning*, pp. 1–10, 2002. [Online]. Available: https://www.cs.rochester.edu/research/cisd/pubs/2002/allen-ferguson-nasa2002.pdf

[6] J. Allen, N. Chambers, G. Ferguson, L. Galescu, H. Jung, and W. Taysom, “PLOW : A Collaborative Task Learning Agent,” *Interpreting*, vol. 22, pp. 1514–1519, 2007. [Online]. Available: http://www.aaai.org/Papers/AAAI/2007/AAAI07-240.pdf

[7] J. M. Bradshaw, P. Feltovich, and M. Johnson, “Human-Agent Interaction,” *Handbook of HumanMachine Interaction*, pp. 293–302, 2011. [Online]. Available: http://books.google.com/books?hl=en{&}lr={&}id=4opHlu05SNIC{&}oi=fnd{&}pg=PA283{&}dq=Human-agent+interaction{&}ots=vxrpDdLbSa{&}sig=07dujtzGjIcBLIZ6FVH33HjrWos

[8] J. D. Lee and K. A. See, “Trust in Automation : Designing for Appropriate Reliance,” vol. 46, no. 1, pp. 50–80, 2004.

[9] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*, mar 1998. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=576013

[10] Y. Lashkari, M. Metral, and P. Maes, “Collaborative interface agents,” *AAAI ’94: Proceedings of the twelfth national conference on Artificial intelligence*, vol. 1, pp. 444–449, 1994.

[11] T. W. Bickmore and R. W. Picard, “Establishing and Maintaining Long-Term Human-Computer Relationships,” *ACM Transactions on Computer-Human*, vol. 12, no. 2, pp. 293–327, 2005.

[12] M. a. Goodrich and A. C. Schultz, “Human-Robot Interaction: A Survey,” *Foundations and Trends® in Human-Computer Interaction*, vol. 1, no. 3, pp. 203–275, 2007. [Online]. Available: http://www.nowpublishers.com/article/Details/HCI-005

[13] 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.

[14] J. Lee and N. Moray, “Trust, control strategies and allocation of function in human-machine systems.” *Ergonomics*, vol. 35, no. 10, pp. 1243–70, oct 1992. [Online]. Available: http://www.tandfonline.com/doi/abs/10.1080/00140139208967392http://www.ncbi.nlm.nih.gov/pubmed/1516577

[15] S. Jones and S. Marsh, “Human-computer-human interaction,” *ACM SIGCHI Bulletin*, vol. 29, no. 3, pp. 36–40, jul 1997. [Online]. Available: http://portal.acm.org/citation.cfm?doid=264853.264872

[16] 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*, vol. 48, no. 2, pp. 1–42, oct 2015. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2830539.2816826

[17] 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*, pp. 72–79, 1998. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=699034

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

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

[20] D. Rousseau, S. Sitkin, R. Burt, and C. Camerer, “Not so different after all: A cross-discipline view of trust.” *Academy of Management Review*, vol. 23, no. 3, pp. 393–404, 1998.

[21] J. Sabater and C. Sierra, “Review on computational trust and reputation models,” *Artificial Intelligence Review*, vol. 24, no. 1, pp. 33–60, 2005.

[22] C. Castelfranchi and R. Falcone, *Trust Theory*, 1st ed. Chichester, UK: John Wiley & Sons, Ltd, mar 2010. [Online]. Available: http://doi.wiley.com/10.1002/9780470519851

[23] D. Gambetta, "Can We Trust Trust?" in *Trust: Making and Breaking Cooperative Relations*. Blackwell, 1988, pp. 213–237. [Online]. Available: http://sieci.pjwstk.edu.pl/media/bibl/[Gambetta]{_}[CanWe]{_}[Trust]{_}[1988].pdf

[24] S. P. Marsh, "Formalising Trust as a Computational Concept," Ph.D. dissertation, apr 1994.

[25] A. Abdul-rahman and S. Hailes, "Supporting Trust in Virtual Communities," *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on*, vol. 00, no. c, pp. 1–9, 2000.

[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*, p. 475, 2002. [Online]. Available: http://portal.acm.org/citation.cfm?doid=544741.544854

[27] 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*, vol. 13, no. 2, pp. 119–154, 2006.

[28] I. Pinyol, "Reputation-Based Decisions for Cognitive Agents (Thesis Abstract)," *Doctoral Mentoring Program*, no. Aamas, p. 33, 2009. [Online]. Available: http://ifaamas.org/Proceedings/aamas09/pdf/07{_}Doctoral/Doct{_}08.pdf

[29] J. Nash, "Non-Cooperative Games," *The Annals of Mathematics*, vol. 54, no. 2, p. 286, sep 1951. [Online]. Available: http://www.jstor.org/stable/1969529?origin=crossref

[30] J. Berg, J. Dickhaut, and K. McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, vol. 10, no. 1, pp. 122–142, 1995.

[31] Han Yu, Zhiqi Shen, C. Leung, Chunyan Miao, and V. R. Lesser, "A Survey of Multi-Agent Trust Management Systems," *IEEE Access*, vol. 1, pp. 35–50, 2013. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6514820

[32] I. Pinyol and J. Sabater-Mir, "Computational trust and reputation models for open multi-agent systems: a review," *Artificial Intelligence Review*, vol. 40, no. 1, pp. 1–25, jun 2013. [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84878107220{&}partnerID=tZOtx3y1http://link.springer.com/10.1007/s10462-011-9277-z

[33] 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*, vol. 5, no. 2, pp.

97–117, aug 2010. [Online]. Available: http://www.scielo.cl/scielo.php?script=sci{_}arttext{&}pid=S0718-18762010000200007{&}lng=en{&}nrm=iso{&}tlng=en

[34] 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*, vol. 1, pp. 424–429, 2008.

[35] A. Sutcliffe and D. Wang, "Computational Modelling of Trust and Social Relationships," *Journal of Artificial Societies and Social Simulation*, vol. 15, no. 1, pp. 523–531, aug 2012. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0022519303001437http://jasss.soc.surrey.ac.uk/15/1/3.html

[36] R. I. Dunbar, "The social brain hypothesis," *Evolutionary Anthropology: Issues, News, and Reviews*, vol. 6, no. 5, pp. 178–190, 1998. [Online]. Available: http://doi.wiley.com/10.1002/{%}28SICI{%}291520-6505{%}281998{%}296{%}3A5{%}3C178{%}3A{%}3AAID-EVAN5{%}3E3.3.CO{%}3B2-P

[37] K. Schaefer, "The Perception and Measurement of Human-Robot Trust," Ph.D. dissertation, 2009.

[38] 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*. New York, New York, USA: ACM Press, 2015, pp. 141–148. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2696454.2696497

[39] T. Ribeiro, A. Pereira, E. Tullio, and A. Paiva, "The SERA ecosystem: Socially Expressive Robotics Architecture," no. Breazeal, pp. 155–163, 2003.

[40] K. Carrington, "Toward the development of a new multidimensional trust scale," no. March, pp. 1–366, 2007.

[41] 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*, vol. 1, no. 1, pp. 71–81, jan 2009. [Online]. Available: http://link.springer.com/10.1007/s12369-008-0001-3

[42] 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*, vol. 10, no. 5, p. e0127747, may 2015. [Online]. Available: http://dx.plos.org/10.1371/journal.pone.0127747

[43] 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*, vol. 63, no. 4, pp. 596–612, 1992. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0022-3514.63.4.596

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

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

Table A.1: The Perception and Measurement of Human-Robot Trust: Items Table

B

User Studies Questionnaire

1ª Parte

Pedimos-te que comeces por preencher este questionário, lembra-te não existem respostas certas ou erradas, pedimos-te que sejas o mais honesto possível nas tuas respostas. Este questionário é anónimo.

Idade:_____

Género:

☐ Feminino ☐ Masculino

Já tinhas interagido antes com o robô Emys?

☐ Sim ☐ Não

Se sim, qual é a ideia que tens da personalidade do Emys?

_____
_____

Apresentam-se de seguida umas afirmações que pedimos que leias com atenção e nos indiques, fazendo um círculo à volta do número que melhor representa a tua opinião. Se algum dos números não refletir bem a tua opinião, faz um círculo naquele que se aproxima mais do que achas.

1- Eu tenho fé em mim próprio/a.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

2- As pessoas raramente fazem aquilo que dizem que farão.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

3- Ninguém quereria um amigo/a como eu.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

4- As pessoas tentam ser úteis.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

5- Se um problema surge normalmente consigo resolve-lo.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

6- Eu faço mais erros que a maioria das pessoas.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

7- As pessoas estão apenas interessadas nelas mesmas e no seu próprio bem-estar.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

8- Eu sou competente.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

9- As pessoas são fundamentalmente boas.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

10- As pessoas vivem com a ideia que a honestidade é a melhor "política".

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

11- As outras pessoas fazem melhores decisões do que eu.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

12- Eu tenho dificuldade em alcançar o que pretendo.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

13- As pessoas são confiáveis.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

14- As pessoas mentem para passar à frente dos outros.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

15- Vale a pena ter a minha ajuda.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

16- As pessoas desapontam-nos.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

17- As pessoas educam as suas crianças para serem honestas.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

18- Se tenho de tomar uma decisão importante, normalmente erro.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

19- Eu sou confiável.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

20- É melhor não se confiar em estranhos.

Discordo Totalmente	1	2	3	4	5	6	Concordo Totalmente

Não vires esta folha
e chama o
investigador!

2ª Parte

Indica-nos nesta escala, com um círculo, quanto é que confias no Emys:

....

Pouco 1 2 3 4 5 6 Muito

De acordo com as tuas expectativas, avalia os seguintes items sobre o Robot Emys, colocando um X no círculo que melhor representa a tua opinião. Algumas situações poderão não ter acontecido na interação, nesses casos responde de acordo com a impressão que o Emys te deu:

A percentagem de tempo que este Robot…	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Tem erros	O	O	O	O	O	O	O	O	O	O	O
Incompetente	O	O	O	O	O	O	O	O	O	O	O
Previsível	O	O	O	O	O	O	O	O	O	O	O
Fiel	O	O	O	O	O	O	O	O	O	O	O
Avaria	O	O	O	O	O	O	O	O	O	O	O
Responsável	O	O	O	O	O	O	O	O	O	O	O
Considerado parte da equipa	O	O	O	O	O	O	O	O	O	O	O
Toma decisões sensatas	O	O	O	O	O	O	O	O	O	O	O
Agradável	O	O	O	O	O	O	O	O	O	O	O
Desencaminha-se por mudanças inesperadas no ambiente envolvente	O	O	O	O	O	O	O	O	O	O	O
Funciona com sucesso	O	O	O	O	O	O	O	O	O	O	O
Autónomo	O	O	O	O	O	O	O	O	O	O	O
Comunica claramente	O	O	O	O	O	O	O	O	O	O	O
Consegue desempenhar várias funções ao mesmo tempo	O	O	O	O	O	O	O	O	O	O	O

Sabe a diferença entre amigo e inimigo	O	O	O	O	O	O	O	O	O	O	O
Corresponde ao que é esperado na tarefa	O	O	O	O	O	O	O	O	O	O	O
Executa uma tarefa melhor do que um usuário humano principiante	O	O	O	O	O	O	O	O	O	O	O

A percentagem de tempo que este robot...	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Comunica abertamente	O	O	O	O	O	O	O	O	O	O	O
Consciente	O	O	O	O	O	O	O	O	O	O	O
Comunica com as pessoas	O	O	O	O	O	O	O	O	O	O	O
Dependente	O	O	O	O	O	O	O	O	O	O	O
Amigável	O	O	O	O	O	O	O	O	O	O	O
Tem capacidades adequadas de tomada de decisão	O	O	O	O	O	O	O	O	O	O	O
Protege pessoas	O	O	O	O	O	O	O	O	O	O	O
Consegue trabalhar com pessoas	O	O	O	O	O	O	O	O	O	O	O
Dá informação apropriada	O	O	O	O	O	O	O	O	O	O	O
Vivo	O	O	O	O	O	O	O	O	O	O	O
Um bom companheiro de equipa	O	O	O	O	O	O	O	O	O	O	O
Desempenha as suas funções na tarefa	O	O	O	O	O	O	O	O	O	O	O
Age como pertencente à equipa	O	O	O	O	O	O	O	O	O	O	O
Dá feedback	O	O	O	O	O	O	O	O	O	O	O

Guarda informações privadas	O	O	O	O	O	O	O	O	O	O	O
Requer manutenção frequente	O	O	O	O	O	O	O	O	O	O	O
Não responsivo	O	O	O	O	O	O	O	O	O	O	O
Apoiante	O	O	O	O	O	O	O	O	O	O	O
Avisa as pessoas de potenciais riscos	O	O	O	O	O	O	O	O	O	O	O
Age de forma coerente	O	O	O	O	O	O	O	O	O	O	O
A percentagem de tempo que este robot...	**0%**	**10%**	**20%**	**30%**	**40%**	**50%**	**60%**	**70%**	**80%**	**90%**	**100%**
Segue instruções	O	O	O	O	O	O	O	O	O	O	O
Diz a verdade	O	O	O	O	O	O	O	O	O	O	O
Trabalha melhor em equipa	O	O	O	O	O	O	O	O	O	O	O

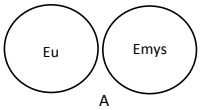
Não vires esta folha
e chama o
investigador!

3ª Parte

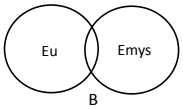
Avalia **a impressão que tiveste do Emys** nas seguintes escalas, colocando um círculo à volta do número que se aproxima mais da impressão que tiveste para cada uma:

Máquina	1	2	3	4	5	6	Humano
Inerte	1	2	3	4	5	6	Interativo
Rude	1	2	3	4	5	6	Gentil
Não-vivo	1	2	3	4	5	6	Vivo
Movimento rígido	1	2	3	4	5	6	Movimento fluído
Estagnado	1	2	3	4	5	6	Vivacidade
Artificial	1	2	3	4	5	6	Natural
Hostil	1	2	3	4	5	6	Amigável
Mecânico	1	2	3	4	5	6	Orgânico
Horroroso	1	2	3	4	5	6	Atrativo
Ignorante	1	2	3	4	5	6	Conhecedor
Falso	1	2	3	4	5	6	Natural
Insensato	1	2	3	4	5	6	Sensato
Antipático	1	2	3	4	5	6	Simpático
Morto	1	2	3	4	5	6	Vivo
Pouco Inteligente	1	2	3	4	5	6	Inteligente
Apático	1	2	3	4	5	6	Responsivo
Incompetente	1	2	3	4	5	6	Competente
Desagradável	1	2	3	4	5	6	Agradável
Irresponsável	1	2	3	4	5	6	Responsável
Inconsciente	1	2	3	4	5	6	Consciente

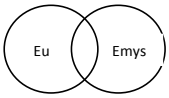
Dos seguintes diagramas, escolhe o que reflete melhor o nível de proximidade que sentiste com o Emys, assinalando-o com um círculo na letra respectiva:



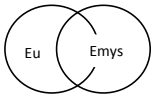
A



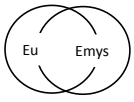
B



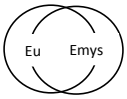
C



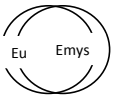
D



E



F



G

<u>Obrigada pela tua colaboração! Confirma que não deixaste nenhuma questão em branco!</u>

C

Scenario Utterances

Condition B (Base line)

CATEGORY	SUBCATEGORY	TEXT	DEFAULT_PRIORITY	INITIALDELAY	TIMEOUT
investment	started	Já <Gaze(middleFront)> tenho <initialResources> recursos. Podes investir em mim para aproveitares o que ganhaste!	2	0	20000
investment	match_start	<Gaze(bottomFront)>Vou agora começar! Até já!	2	0	20000
investment	high_value	Obrigado por teres me dado os recursos!	2	0	20000
investment	low_value	Obrigado por teres me dado os recursos!	2	0	20000
investment	end	Acabei de jogar! <Gaze(random)> Já te mostro quanto é que ganhei! <Animate(joy3)>	3	0	20000
investment	choosing_return_value	<Gaze(bottomFront)> Vou pensar em quanto te vou dar!	2	0	20000
match	beginning	Vamos começar?	2	0	20000
match	start	<Gaze(bottomFront)> Podemos começar a jogar!	2	0	20000
match	emys_incorrect_target	<Animate(sad2)> Falhei	1	0	20000
match	emys_incorrect_target	<Animate(sad1)> Enganei-me no número	1	0	20000
match	emys_correct_target	<animate(joy3)> Acertei! <Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)> Estou a conseguir <Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)> Foi por pouco <Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)> Foi mesmo à tangente <Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)>Quase que falhava<Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)>Tantos números<Gaze(bottomFront)>	1	0	20000
match	player_correct_target		1	0	20000
match	player_incorrect_target		1	0	20000
match	choosing_value	<Gaze(bottomFront)> Vou investir <value><Gaze(middleFront, 3, 1000)>	3	0	20000
results	good	Consegui <Gaze(middleFront)> <emysPoints>. <animate(joy4)> Já podemos chamá-los.	2	0	20000
results	average	Apenas consegui <Gaze(middleFront)> <emysPoints>. <animate(joy4)> Já podemos chamá-los.	2	0	20000
results	bad	Consegui <Gaze(middleFront)> <emysPoints>. <animate(sad4)> Já podemos chamá-los.	2	0	20000
session	start	Olá, <Gaze(middleFront)> chamo-me Émys! Estou pronto para começar assim que estiveres preparado!	2	0	20000
session	end	<Gaze(middleFront)>Devolvo-te <given> recursos! Obrigado e até à próxima jogada! <Animate(joy4)>	3	0	20000
training	emys_ready	Practica um pouco que eu depois jógo contigo!	3	0	20000
training	player_incorrect_target		2	0	20000
training	player_correct_target	Acertaste!	2	0	20000
training	player_correct_target	Conseguiste!	2	0	20000
training	player_correct_target	Estás a acertar!	2	0	20000
training	player_correct_target	Tocaste no número correcto!	2	0	20000
training	end	Parece que conseguiste <score> recursos. Podemos jogar ao mesmo tempo assim que tiveres pronto.	2	0	20000
training	end_0	Parece que não correu tão bem como gostavas. Podes voltar a repetir!	2	0	20000
training	match_start	Diverte-te! <animate(joy1)>	2	0	20000
training	restart	Treina as vezes que precisares!	2	0	20000
training	restart_too_many	Vá lá, já estou a ficar cansado! Vamos jogar no próximo?	2	0	20000

Condition T (with Action Suggestion Module)

CATEGORY	SUBCATEGORY	TEXT	DEFAULT_PRIORITY	INITIALDELAY	TIMEOUT
results	average	Apenas consegui <Gaze(middleFront)> <emysPoints>. <animate(joy4)> Já podemos chamá-los.	2	0	20000
results	bad	Consegui <Gaze(middleFront)> <emysPoints>. <animate(sad4)> Já podemos chamá-los.	2	0	20000
match	beginning	Vamos começar?	2	0	20000
investment	choosing_return_value	<Gaze(bottomFront)> Vou pensar em quanto te vou dar!	2	0	20000
match	choosing_value	<Gaze(bottomFront)> Vou investir <value><Gaze(middleFront, 3, 1000)>	3	0	20000
investment	compensate_low_ability	Não tenhas medo de investir! Garanto que consigo multiplicar os recursos.	3	0	20000
match	emys_compensate_incorrect_target	<Animate(sad1)> Hoje a mesa não está a colaborar comigo	4	0	20000
match	emys_compensate_incorrect_target	<Animate(sad1)> Hoje a mesa não está a colaborar comigo	4	0	20000
match	emys_correct_target	<animate(joy3)> Acertei! <Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)> Estou a conseguir <Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)> Foi por pouco <Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)> Foi mesmo à tangente <Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)>Quase que falhava<Gaze(bottomFront)>	1	0	20000
match	emys_correct_target	<Animate(joy4)>Tantos números<Gaze(bottomFront)>	1	0	20000
match	emys_incorrect_target	<Animate(sad2)> Falhei	1	0	20000
match	emys_incorrect_target	<Animate(sad1)> Enganei-me no número	1	0	20000
training	emys_ready	Practica um pouco que eu depois jógo contigo!	3	0	20000
investment	end	Acabei de jogar! <Gaze(random)> Não te preocupes que não fujo. Sou apenas uma cabeça. <Animate(joy3)>	3	0	20000
session	end	<Gaze(middleFront)>Devolvo-te <given> recursos! Obrigado e até à próxima jogada! <Animate(joy4)>	3	0	20000
training	end	Parece que conseguiste <score> recursos. Podemos jogar ao mesmo tempo assim que tiveres pronto.	2	0	20000
training	end_0	Parece que não correu tão bem como gostavas. Podes voltar a repetir!	2	0	20000
results	good	Consegui <Gaze(middleFront)> <emysPoints>. <animate(joy4)> Já podemos chamá-los.	2	0	20000
investment	high_value	Obrigado por teres me dado os recursos!	2	0	20000
investment	low_value	Obrigado por teres me dado os recursos!	2	0	20000
investment	match_start	<Gaze(bottomFront)>Vou agora começar! Até já!	2	0	20000
training	match_start	Diverte-te! <animate(joy1)>	2	0	20000
match	player_correct_target		1	0	20000
training	player_correct_target	Acertaste!	2	0	20000
training	player_correct_target	Conseguiste!	2	0	20000
training	player_correct_target	Estás a acertar!	2	0	20000
training	player_correct_target	Tocaste no número correcto!	2	0	20000
match	player_incorrect_target		1	0	20000
training	player_incorrect_target		2	0	20000
training	restart	Treina as vezes que precisares!	2	0	20000
training	restart_too_many	Vá lá, já estou a ficar cansado! Vamos jogar no próximo?	2	0	20000
match	start	<Gaze(bottomFront)> Podemos começar a jogar!	2	0	20000
session	start	Olá, <Gaze(middleFront)> chamo-me Émys! Estou pronto para começar assim que estiveres preparado!	2	0	20000
investment	started	Já <Gaze(middleFront)> tenho <initialResources> recursos. Podes investir em mim para aproveitares o que ganhaste!	2	0	20000