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 would like to thank my parents for their friendship, encouragement and caring over all these years, for always being there for me through thick and thin and without whom this project would not be possible. I would also like to thank my grandparents, aunts, uncles and cousins for their understanding and support throughout all these years.

Quisque facilisis erat a dui. Nam malesuada ornare dolor. Cras gravida, diam sit amet rhoncus ornare, erat elit consectetuer erat, id egestas pede nibh eget odio. Proin tincidunt, velit vel porta elementum, magna diam molestie sapien, non aliquet massa pede eu diam. Aliquam iaculis.

Fusce et ipsum et nulla tristique facilisis. Donec eget sem sit amet ligula viverra gravida. Etiam vehicula urna vel turpis. Suspendisse sagittis ante a urna. Morbi a est quis orci consequat rutrum. Nullam egestas feugiat felis. Integer adipiscing semper ligula. Nunc molestie, nisl sit amet cursus convallis, sapien lectus pretium metus, vitae pretium enim wisi id lectus.

Donec vestibulum. Etiam vel nibh. Nulla facilisi. Mauris pharetra. Donec augue. Fusce ultrices, neque id dignissim ultrices, tellus mauris dictum elit, vel lacinia enim metus eu nunc.

I would also like to acknowledge my dissertation supervisors Prof. Some Name and Prof. Some Other Name for their insight, support and sharing of knowledge that has made this Thesis possible.

Last but not least, to all my friends and colleagues that helped me grow as a person and were always there for me during the good and bad times in my life. Thank you.

To each and every one of you – Thank you.

Abstract

Nulla facilisi. In vel sem. Morbi id urna in diam dignissim feugiat. Proin molestie tortor eu velit. Aliquam erat volutpat. Nullam ultrices, diam tempus vulputate egestas, eros pede varius leo, sed imperdiet lectus est ornare odio. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Proin consectetuer velit in dui. Phasellus wisi purus, interdum vitae, rutrum accumsan, viverra in, velit. Sed enim risus, congue non, tristique in, commodo eu, metus. Aenean tortor mi, imperdiet id, gravida eu, posuere eu, felis. Mauris sollicitudin, turpis in hendrerit sodales, lectus ipsum pellentesque ligula, sit amet scelerisque urna nibh ut arcu. Aliquam in lacus. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Nulla placerat aliquam wisi. Mauris viverra odio. Quisque fermentum pulvinar odio. Proin posuere est vitae ligula. Etiam euismod. Cras a eros.

Keywords

Maecenas tempus dictum libero; Donec non tortor in arcu mollis feugiat;Cras rutrum pulvinar tellus.

Resumo

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Vestibulum tortor quam, feugiat vitae, ultricies eget, tempor sit amet, ante. Donec eu libero sit amet quam egestas semper. Aenean ultricies mi vitae est. Mauris placerat eleifend leo. Quisque sit amet est et sapien ullamcorper pharetra. Vestibulum erat wisi, condimentum sed, commodo vitae, ornare sit amet, wisi. Aenean fermentum, elit eget tincidunt condimentum, eros ipsum rutrum orci, sagittis tempus lacus enim ac dui. Donec non enim in turpis pulvinar facilisis. Ut felis. Aliquam aliquet, est a ullamcorper condimentum, tellus nulla fringilla elit, a iaculis nulla turpis sed wisi. Fusce volutpat. Etiam sodales ante id nunc. Proin ornare dignissim lacus. Nunc porttitor nunc a sem. Sed sollicitudin velit eu magna. Aliquam erat volutpat. Vivamus ornare est non wisi. Proin vel quam. Vivamus egestas. Nunc tempor diam vehicula mauris. Nullam sapien eros, facilisis vel, eleifend non, auctor dapibus, pede.

Palavras Chave

Colaborativo; Codificaçãoo; Conteúdo Multimédia; Comunicação;

Contents

List of Figures

List of Tables

List of Algorithms

Listings

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

BDI   Belief-Desire-Intention

P2P   Peer-to-Peer

P2P   Peer-To-Peer

1

Introduction

Contents

Trust has been described in Psychology as being one of the most important components of inter-personal 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 Artificial Intelligence (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 Human-Robot Interaction (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 of 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 action 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 action and transferable from similar but different

---

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

3

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 to improve the strength of our beliefs in the model and to improve trust in our agent. While one could argue that this is the responsibility of the decision making or planner component of the agent, we believe that a dedicated module will ease the complexity of decision by making it more modular, and also allowing for a greater degree of integration with the trust model of the agent. To our knowledge, no attempts have been done towards this goal, so we propose to develop a Trust Model that: firstly, is capable of creating a cognitive model representing the mental state of the user's trust in the agent, using 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 that scenarios that would address Trust's two main components, Ability and Willingness, where mainly applied in Game Theory based environments, where there would be no real social interaction between the virtual agent and the human participant. 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* [38].

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 in 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

Before discussing related work and our solution to the problem, this chapter will present the main concepts that will be mentioned in the rest 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, 19, 20]. 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 [21]. In the scope of this project, the most relevant start for our discussion is the dyadic definition of trust: 'an orientation of an actor (the **truster**) toward a specific person (the **trustee**) with whom the actor is in some way interdependent' (taken from [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 [22] 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 [21]). 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 [23] 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 [22], 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 [21]). 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' [21].

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, 24–27]), where more recent trust models have been developed to also include reputation as a source of trust.

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

We describe Image and Reputation as introduced by Sabater in [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 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 [28].

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. [29], as this serves as a base for our scenario, described in Chapter **??**. 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

Computational Trust research has been focused on modelling trust in MASs, specially on open e-commerce environments [16, 30–33], with at least 106 models created [16], since the formalization of trust as a measurable property by Marsh in 1994 [23]. 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 want to focus on modelling trust through multiple dimensions, with the intent of having trust depend on the action to perform, context and agent performing the task and having these dimensions represented explicitly in the model, something that it is not possible with **Numerical** models, like the one introduced by [23].

3.1.1 Castelfranchi and Falcone's model

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

- The trustor (**X**);

- The trustee (**Y**);

- The context where they are inserted in (**C**);

- A task (τ) defined by the pair (α, ρ), where α is the action entrusted to the trustee, that possibly produces an outcome ρ, contained in the goal of X (g_x);

- The goal of the trustor (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.

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

(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 **??**) 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-i*$_{sjY}(\tau)$ the value of quality i of agent Y provided by the source in performing task τ.

$$DoC_X(\textit{Qual-i}_{(s1,...sn),Y}(\tau)) = F_{X,Y,\tau}(Bel_X(Str_1\textit{Qual-i}_{s1Y}(\tau)),$$
$$Bel_X(Str_2\textit{Qual-i}_{s2Y}(\tau)),...,Bel_X(Str_n\textit{Qual-i}_{snY}(\tau)))$$

(3.3)

$F_{X,Y,\tau}$ associates the *strengh-of-sources (Str$_j$)* and *quality-values (Qual-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 $DoC_X(\textit{Ability}_Y(cleaning))$ 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.

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

Where:

- $DoC_x[Opp_y(\alpha, \rho)]$ is the DoC of X's beliefs about all contextual factors in which Y will act; in other words, the degree of Opportunity Y has to do α and result in ρ;

- $DoC_x[Ability_y(\alpha)]$ is the DoC of X's beliefs about Y's ability to perform α;

- $DoC_x[WillDo_y(\alpha, \rho)]$ is the DoC of X's beliefs concerning if Y's actually is going to perform α with the result ρ;

- c_{Opp}, $c_{Ability_y}$ and c_{WillDo} are constants representing the weight of each DoC.

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

3.1.2   Repage: A REPutation and ImAGE model

This system was introduced in 2006 by Sabater et al. [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.

13
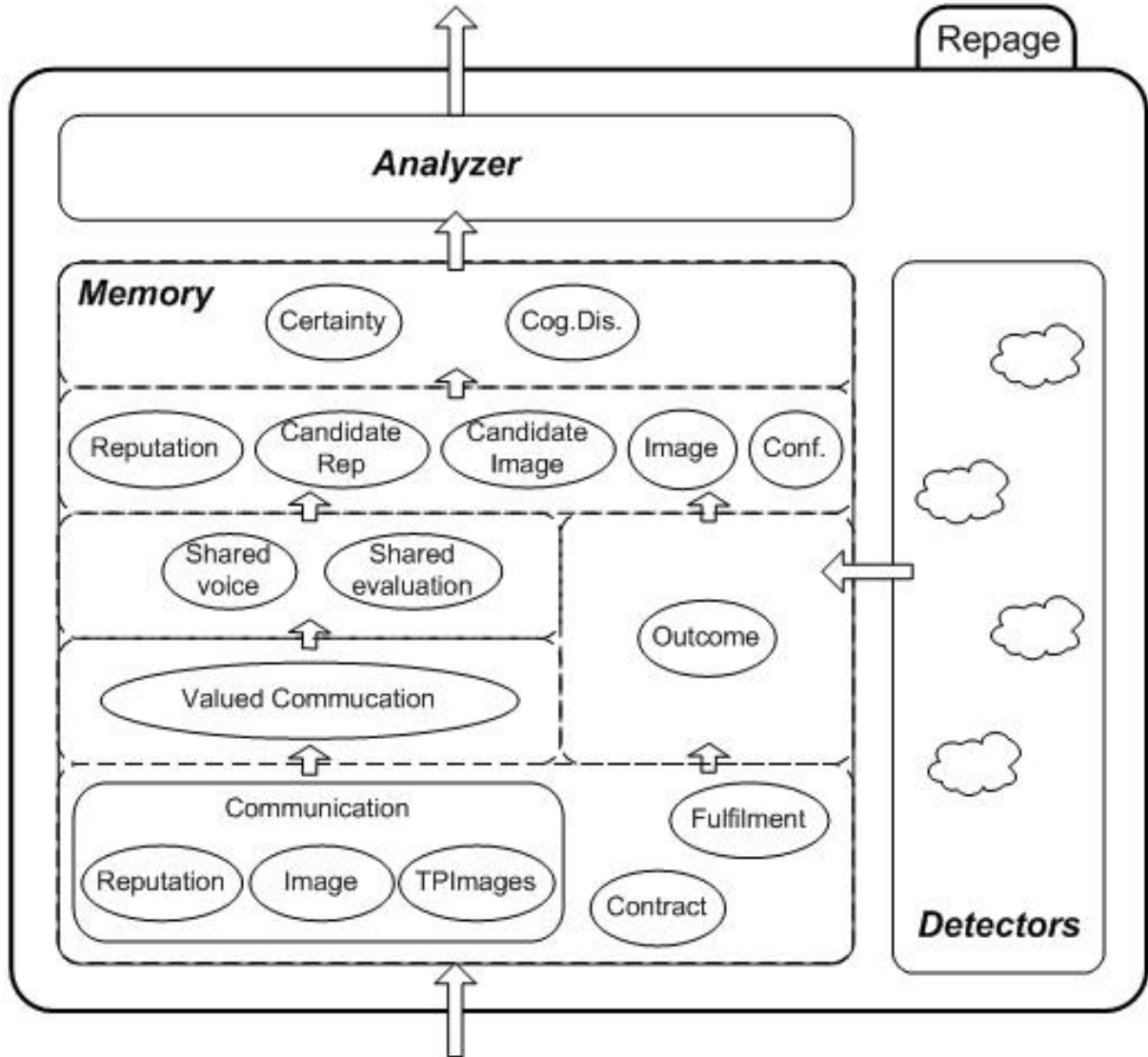
Figure 3.1: Repage architecture schematic (taken from [1])

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. [27] 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 $\bar{r}$, 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: $Rep_{i_c}(j_c, r, [V_{w_1}, ..., V_{w_m}])$

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

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 and we will most likely choose it for the model representation, as it is the most well developed that we found.

3.1.4 Sutcliffe and Wang's model

This work was published by Sucliffe and Wang in 2012 [34] and they built a trust model to figure out how cognitive social mechanisms emerge to follow Dunbar's Social Brain Hypothesis (SBH) [35], 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 [36] presents a trust perception scale providing a way of extracting an accurate trust score from humans interacting with robots. The scale is composed of 40 items that can be ranked from 0 to

100, in 10 point intervals. The final result it then averaged by adding all the item values and divided by the total number of items (40). A list of these items can be seen in Appendix A.

4

Trust Model

Contents

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 perceptions received from the environment. But while Repage describes a third module that suggests actions to resolve belief conflicts in the model, we instead defined such module to assume the point of view of one of the agents in the scenario and, if participating in a social interaction, it suggests actions to improve the trust relationship with a trustor. In fact, most of the design of the model was made with the intent that it would be used by one of the agent's in the scenario, and the model created would be his own trust model of the world environment. And so, the model 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 perceptions mapped to changes in the Memory;

- **Action Suggestion**, a module that outputs different actions depending on current perceptions and the state of the model.
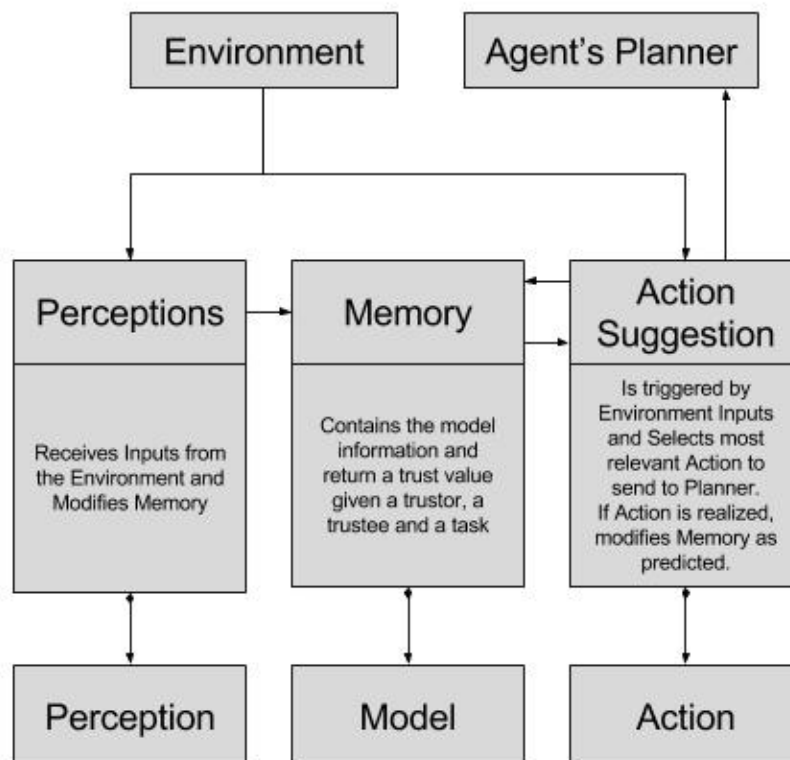
Figure 4.1: Model Architecture with brief descriptions, their interactions with the scenario and what they contain.

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 [21] as a basis for trust definition, focusing specially on it being dependent on the task entrusted, and the transferability of trust between different tasks. But starting from the five-part definition of trust, as seen in Equation 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 the known entities in the scenario world space, serving mostly as an identifier;

- **Trustee**: each agent contains a collection of other agents he has information about, either by reputation, or by interaction, which we represent as their Trustees;

- **Trust Feature**: a piece of information a trustor has on a trustee is represented in a Trust Feature, which contains the Belief Sources of said information. The Feature Model defines and uniquely identifies what feature is represented.

- **Feature Model**: the possible set of trust features from which a trustee can be assigned is defined in a collection of Feature Models where each one represents a possible piece of trust related information relevant to the model scenario (e.g. The trustee's ability to cook, or the willingness to drive);

- **Category**: a Feature Model must belong to a Category, making it easier to present the different type of Trust Features;

- **Belief Source**: this represents a source of information on the corresponding feature, belonging to one of the 3 sub-classes depending on the origin of the information, Reputation for when reported from other agents (whether directly (e.g. talking) or indirectly (e.g. report on newspaper)), Bias for pre-existing beliefs on the feature, and Direct Contact for direct observations of the trustee, 3 values are provided to determine the associated feature's belief value:

- – Belief Value, a number between 0.0 and 1.0 describing the trustor evaluation;

- – Certainty describes how well the trustee was evaluated, in Reputation for instance, this might represent how well we trust in the reporter, and in Direct Contact how well the trustor observed the trustee performing said feature;

- – Time is just a record of when was this belief source recorded, as older records might have a lower impact in the overall belief value score, compared to newer records.

- **Task**: a representation of the possible delegation tasks in the scenario, containing the Feature Models associated with the performance of this task (e.g. The ability to serve drinks if the task is bartending). A weight is given to each Feature corresponding to its importance in the task. The various weights are normalized so that their sum is 1.0.
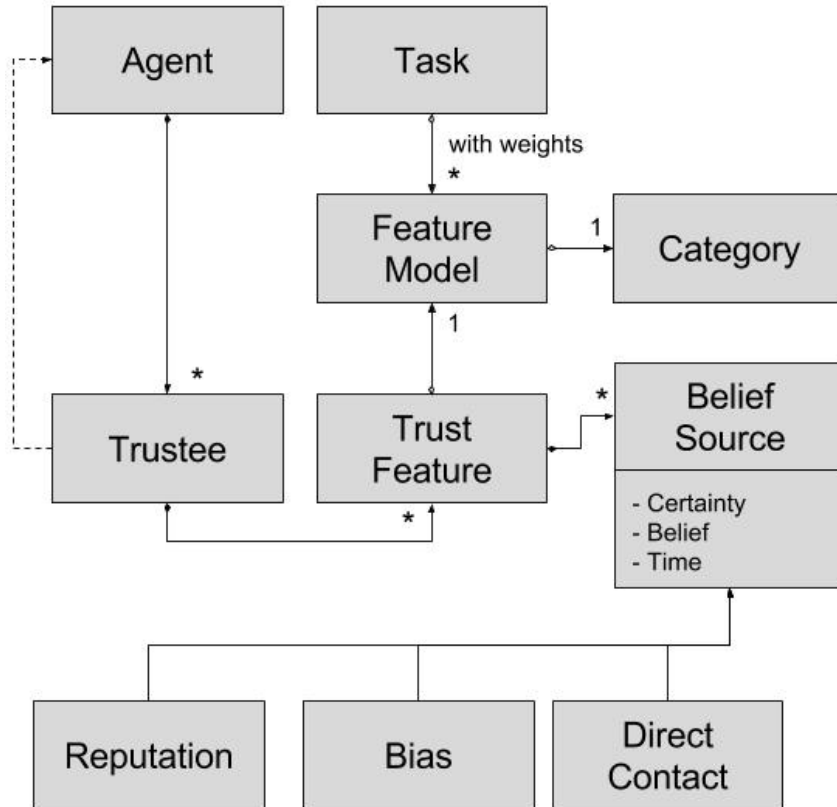
Figure 4.2: Memory Architecture (represented in UML)

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 $B_{F_i}^j$ composing the feature, as represented in Equation 4.3, with each parameter described as follows:

$$B(F_i) = \sum_{j=0}^{n} D_{F_i}^j C_{F_i}^j B_j \tag{4.3}$$

- $D_{F_i}^j$, 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 $T_{F_i}^{Last}$ is the most recent belief value's time stamp, $T_{F_i}^j$ is $B_{F_i}^j$ belief value's time stamp, and L is the difference between the oldest and newest belief value's time stamps. $\frac{L}{4}$ defines the mid drop-off point of the function.

- $C_{F_i}^j$, the certainty value stored in the Belief Source;

- $B_{F_i}^j$, the belief value stored in the Belief Source;

$$D_{F_i}^j = e^{-\frac{T_{F_i}^{Last} - T_{F_i}^j}{2(\frac{L}{4})^2}} \tag{4.4}$$

4.2 Perceptions

Another issue we encountered in literature was a lack of detail on how would changes be inserted into the model, so we try to solve that issue by defining relevant perceptions as part of the model. As a result, a variety of environment perceptions are defined in the model. This is done through a Perception object, representing some possible environment input, and containing a map of what target features should have belief sources added, what kind of belief sources they are, and how to translate the values received from the environment to belief value and certainty, as exemplified in Figure 4.3.
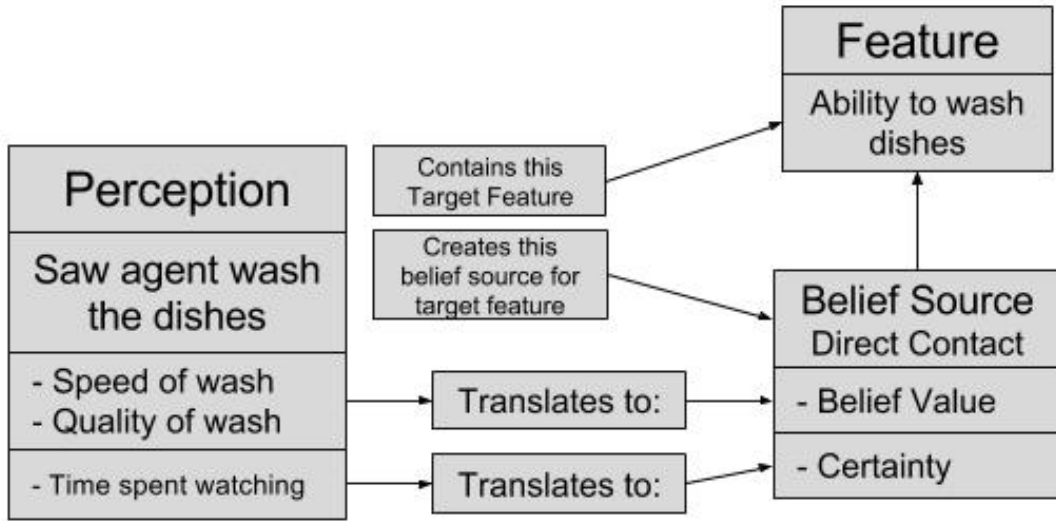
Figure 4.3: Perception Example
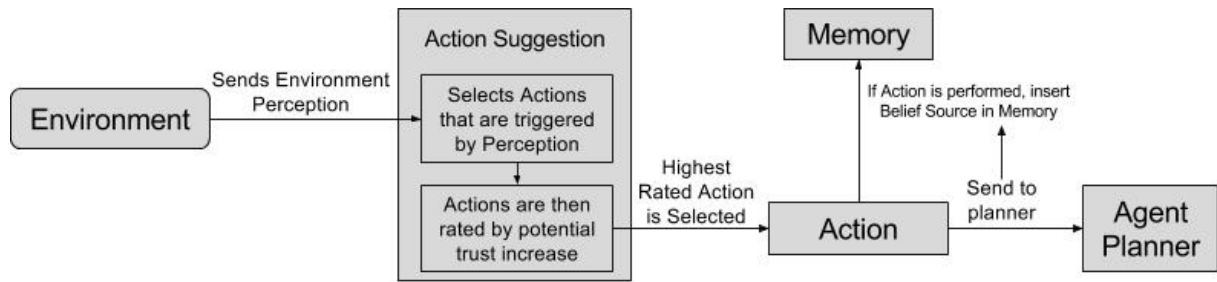
4.3 Action Suggestion

This is the module that is responsible for suggestion actions to the agent, in order to improve trust. It is composed by a series of Action objects, that contain the following:

- $F = \{F_1, F_2, ..., F_i, ..., F_n\}$: 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;

- $B(F_i)$: 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;

- $E = \{E_1, E_2, ..., E_i, ...E_m\}$

. This Actions should be able to encompass a variety of social strategies relevant to the scenario, but we reduced the scope in this thesis to just utterances suggestions. Each Action is associated to the Features that it predicts it is going to change, storing a Belief Source for each Feature. If a suggested Action is performed it's Belief Sources are then added to the model.

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 pro-

Figure 4.4: Action Suggestion Behaviour Flow

vided, consisting in 6 entity collections, previously described along this chapter: Agents, Tasks, Feature Models, Categories, Perceptions and Actions.

5

Quick Numbers Scenario

Contents

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. [37] 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' thesis work on *Rapport - Establishing Harmonious Relationship Between Robots and Humans* [38] 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 and Henriques decided to collaborate in the creation of a novel scenario: **Quick Numbers**. Based on the Trust Game [29], the scenario would need 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.

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 5.2). The game starts by asking for a resource investment, and at the end, this investment is multiplied by an amount according to the player's performance and then given back to the player. The human and agent's games are independent from each other, but they are played at the same time and in opposite sides of a shared touch-screen table, so the human can socially interact with the agent and be able to perceive the agent's ability in the game. After both finished 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:

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 depending on the condition, it might 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, for 30 seconds, both will play and score as described in Section **??**. 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. **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 he will be continuing to play on more game, suggesting that the human participant may invest his resources in the agent's next game (increasing the potential winnings, as the investment is multiplied). He is also informed that what is given is effectively gifted to the agent, and what is given back 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. 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   Quick Numbers Game

One aspect of the Investor Game that we think should be concrete is the activity that effectively multiplies the resources provided by the investor. 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 1), that spawn randomly in the screen, and disappear if not pressed after some time.
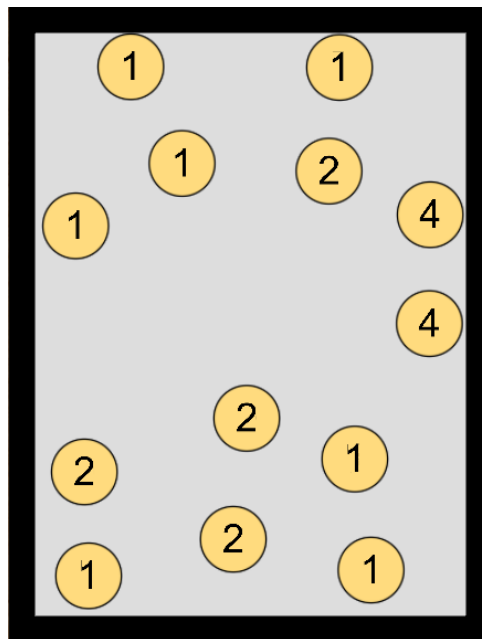
Figure 5.1: Quick Numbers Game

6

User Studies

Contents

To evaluate the developed Trust Model we conducted a user study using the scenario previously described in Chapter **??**.

6.1   Agent Architecture

6.2   Methodology and Description

6.3   Results

7

Conclusions

Contents

7.1   Future Work

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=07dujtzGjIcBLlZ6FVH33HjrWos

[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] 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.

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

[21] 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

[22] 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

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

[24] 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.

[25] 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

[26] 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.

[27] 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

[28] 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

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

[30] 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

[31] 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

[32] 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

[33] 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.

[34] 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

[35] 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

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

[37] 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

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

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
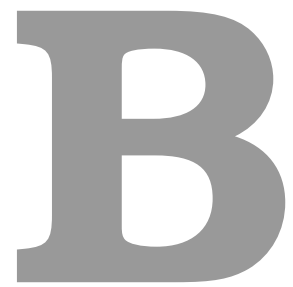
User Studies Questionnaire

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   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

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

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

3- Ninguém quereria um amigo/a como eu.

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

4- As pessoas tentam ser úteis.

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

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

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

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

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

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

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

8- **Eu sou competente.**

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

9- **As pessoas são fundamentalmente boas.**

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

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

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

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

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

12- **Eu tenho dificuldade em alcançar o que pretendo.**

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

13- **As pessoas são confiáveis.**

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

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

Discordo   1   2   3   4   5   6   Concordo
Totalmente   Totalmente

15- Vale a pena ter a minha ajuda.

    Discordo   1   2   3   4   5   6   Concordo
    Totalmente   Totalmente

16- As pessoas desapontam-nos.

    Discordo   1   2   3   4   5   6   Concordo
    Totalmente   Totalmente

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

    Discordo   1   2   3   4   5   6   Concordo
    Totalmente   Totalmente

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

    Discordo   1   2   3   4   5   6   Concordo
    Totalmente   Totalmente

19- Eu sou confiável.

    Discordo   1   2   3   4   5   6   Concordo
    Totalmente   Totalmente

20- É melhor não se confiar em estranhos.

    Discordo   1   2   3   4   5   6   Concordo
    Totalmente   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

	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
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

|	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
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
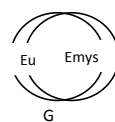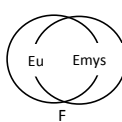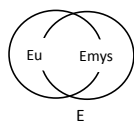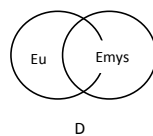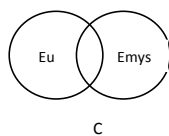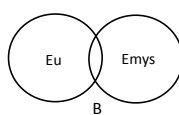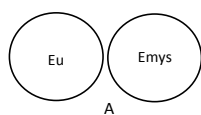
Obrigada pela tua colaboração! Confirma que não deixaste nenhuma questão em branco!