Trust in Human and Autonomous Agent Interaction

Nuno Xu
nuno.xu@tecnico.ulisboa.pt

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

Abstract What thesis is about and contributions

Key words: rapport, Human Robot Interaction (HRI)

Contents

1 Introduction

Trust has been described in Psychology as being one of the most important components of interpersonal relationships [1]. It is undeniable the need of Trust to promote cooperation and collaboration between two parties, either in deciding who should one collaborate with, or even on what we should the other party with. 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 [2].

As Artificial Intelligence (AI) Research gravitates towards the development of Intelligent Agent Systems [3], out of which 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 focus on Human Agent Interaction (HAI). While the amount of literature has been increasing, we found surprising that not enough work has been done in HAI focusing on Trust, other than on design issues [8] and the sub-field of HRI [9, 10], specially when so much has been done regarding Trust in Automation [2, 11, 12] and in modelling Trust and Reputation in Multi-Agent System (MAS) [13].

Reeves and Nass have shown that people apply social rules to Human Computer Interaction (HCI), and this can logically be extended to HAI [14]. So as autonomous agents evolve to better perform collaborative tasks with humans, which demands some amount of social interaction, the active agent must seek out to improve the trust relationship it has with the user [15]. To that goal, we propose to develop two self-contained agent modules, firstly, capable of creating a cognitive model representing the mental state of the user's trust in the agent, and secondly, provide input on what actions should be used to improve or worsen trust on the agent. We will ascertain this project's objectives by integrating the module in an agent implementation, currently finishing development, that is capable of acting as one of the players in the *Split or Steal* scenario, introduced in the British game show *Golden Balls* [16], the scenario is further described in Section 5.

We hope that this project will allow, not only to ensure proper reliance in the agent's task [2], but to also provide additional insight on how certain actions affect human trust on agents differently from one another.

1.1 Goals

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

- Create a cognitive trust model capable of representing human trust towards the agent;
- Develop a decision making module that aims to rank actions as positively or negatively affecting trust in the agent;
- Study what factors most influence trust in certain interactive actions.

In Section 2 we will present a brief summary of main concepts used in this project. Section 3 will discuss the trust models already developed in the field of

MAS. Then in Section 4 a description of our solution architecture will be presented. Finally in Section 5 we will describe how we will evaluate the objectives.

2 Background

Before discussing related work and our solution to the problem, we will present the main concepts that will be mentioned in the rest of this report, specifically regarding Trust, Reputation and Game Theory.

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. So it has been studied in a multitude of disciplines, from Psychology and Sociology, to Philosophy and Economy [12, 17, 18]. 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 [19]. But in the scope of this project, the most relevant start for our discussion is the dyadic definition of trust, where it is defined as: 'an orientation of an actor (the truster) toward a specific person (the trustee) with whom the actor is in some way interdependent' (taken from [1]), as we want to focus on interpersonal relationships. This definition has been expanded throughout the literature, often adapted to fit the context or scope of the work, but three main definitions were highlighted:

- First, Gambetta [20] 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 [19]). This is accepted by most authors as one of the most classical definitions of trust, but we agree with [19], in that it is restrictive it's uni-dimensionality, as it only refers to predictability of the trustor, and does not take into account competence in executing the given action.

- Marsh [21] was the first author to formalize trust as a measurable Computational Concept, continuing the perspective of reducing trust to a numerical value, set by Gambetta [20], 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 [19]). In this definition our opinions also match with [19], regarding that it 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 defined a new definition and paradigm for Computational Trust, introducing a Cognitive aspect to it [22]. 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 hold my sword, but not to swing it.).

2.1.1 Castelfranchi and Falcone's Trust More explicitly, Castelfranchi and Falcone [22] state that Trust can be defined with a central core, composed by a five-part relation, between the trustor, the trustee, the context where they are inserted in, the action and outcome done by the trustee, and the goal of the trustor. 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 the trustee is being delegated for, 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. It's important to note, that following this definition, Trust must imply that the trustor is taking some kind of risk by delegating a task to the trustee. Be it because the trustee may not able to perform the task, or that he may purposely ruin the task and go against the trustor goals.

2.2 Reputation

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 [23–27]), where more recent Trust models have been developed to also include reputation as a source of Trust, where the agent is not influenced only by the image one has of the subject, but also by what other agents say about it.

For the purpose of this report, reputation of an agent is defined as the combined trust opinion that the other agents that have manifested or provided about a particular subject. This reputation can and should be different from the individual image that various agents may have about the subject.

2.3 Game Theory

Game Theory is the field of study that defines and analyses situations involving conflict or cooperation between multiple intelligent decision makers. These situations are called a game, and they are distilled to their core argument, by defining the limited and simple set of actions that the players may perform, and how do they affect the players. It then analyses the decision strategies for each player, by assuming that both will try to maximise their payoff (how much the player gains) with their action. To better explain the concepts we want to present, we will introduce one of the most common exemplary models of Game Theory, the Prisoner's Dilemma.

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

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

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

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

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

Table 1: Prisoner's Dilemma Payoff Matrix

In the game we can say that *Defecting* **dominates** *Cooperating*, as for any action that the adversary player may choose, *Defecting* always gives a better payoff for the individual player [28].

3 Related Work

Computational Trust research has been focused on modelling trust in MASs, specially on open e-commerce environments [13, 29–32], with at least 106 models created [13], since the formalization of trust as a measurable property by Marsh in 1994 [21].

3.1 Trust Models

For related work concerning Trust Models we will focus on Cognitive Trust Models, first introduced by Castelfranchi et. al. [22], as we want to focus on modelling trust based on the Agent's mental state, specially when considering that our goal is to apply the trust model to human partners.

3.1.1 Castelfranchi et. al.

Having introduced the concept of Cognitive Trust Models, this author's model is generally used as a basis for most other authors,

3.1.2 Repage

3.1.3 BDI + Repage

3.2 The Perception and Measurement of Human-Robot Trust

Schaefer [33] presents a trust perception scale, that provides a way of extracting an accurate trust score from humans interacting with robots. The scale is composed of 40 items that can be ranked from 0 to 100, in 10 point intervals. The final result it then averaged by adding all the item values and divided by the total number of items (40).

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

3.3 Discussion

Based on everything and what everyone is doing or done, these are the problems, these are the advantageous and disavantageous of the aproach.

4 Proposed Solution

- Architecture/Model: does not need to be very detailed but enough to understand what it is trying to be done. Tools being used, etc
- Evaluation: How the work will be evaluated.
- Planning (GANTT maybe)

Aproach - Makes sense because you talked previously

As [paper do iterative perceptual learning], using hand crafted rules [7, 12] despite being intuitive use very shallow features and the development of these rules is not trivial. E.g.

4.1 Subsection Heading Here

Subsection text here. A figure can be inserted like the example of Figure **??**.

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

Table 2: My caption

5 Evaluation

As stated in Section 1, the project's evaluation will be done by integrating the developed modules in a agent, now finishing development, that is capable of acting as a player in the *Split or Steal* scenario, introduced in the British television show *Golden Balls* [16]. The scenario involves two players and stands as follows:

- A large sum of money is prized for the game;
- Each player receives two balls, one has 'Steal' written inside while the other has 'Split';
- The balls can be stealthily open by the players, so only each player knows is written in what ball;
- The players then have some time to discuss and negotiate between one another;
- Finally the players choose one of their balls and show their content simultaneously and the game finalizes, giving one of the following results:
  - If both players choose 'Split', they split the prize money evenly;
  - If one picks 'Steal' and the other 'Split', the stealer gets all the prize money;
  - If both pick 'Steal', they both lose all the prize money.

From a game theory standpoint, this scenario is a variation on *Prisoner's Dilemma*, described in Section 2.3.1, where defecting only weakly dominates cooperating, because if an opposing player picks 'Steal' we get nothing whether we pick 'Steal' or 'Split', so 'Steal' only dominates if the opposing player picks 'Split'.

The negotiation phase of the game is the one we are going to focus on and it is going to involve

6 Conclusion

The conclusion goes here. This is more of the conclusion.

Acknowledgment

The author would like to thank...

References

1. Simpson, J.a.: Foundations of interpersonal trust. In: Social psychology: Handbook of basic principles (2nd ed.). (2007) 587–607
2. Lee, J.D., See, K.A., City, I.: Trust in Automation : Designing for Appropriate Reliance. **46**(1) (2004) 50–80
3. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edition. (2009)

4. Grosz, B.J.: Collaborative Systems. AI Magazine (1996) 67–85
5. Allen, J., Ferguson, G.: Human-machine collaborative planning. International NASA Workshop on Planning (2002) 1–10
6. Allen, J., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Taysom, W.: PLOW : A Collaborative Task Learning Agent. Interpreting **22** (2007) 1514–1519
7. Bradshaw, J.M., Feltovich, P., Johnson, M.: Human-Agent Interaction. Handbook of HumanMachine Interaction (2011) 293–302
8. Bickmore, T.W., Picard, R.W.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer-Human **12**(2) (2005) 293–327
9. Goodrich, M.a., Schultz, A.C.: Human-Robot Interaction: A Survey. Foundations and Trends® in Human-Computer Interaction **1**(3) (2007) 203–275
10. van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D.H.J., Haselager, W.F.G.: Do Robot Performance and Behavioral Style affect Human Trust ? International Journal of Social Robotics (2014)
11. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**(10) (oct 1992) 1243–70
12. Jones, S., Marsh, S.: Human-computer-human interaction. ACM SIGCHI Bulletin **29**(3) (jul 1997) 36–40
13. Granatyr, J., Botelho, V., Lessing, O.R., Scalabrin, E.E., Barthès, J.P., Enembreck, F.: Trust and Reputation Models for Multiagent Systems. ACM Computing Surveys **48**(2) (oct 2015) 1–42
14. Reeves, B., Nass, C.: The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places. (mar 1998)
15. Lashkari, Y., Metral, M., Maes, P.: Collaborative interface agents. AAAI '94: Proceedings of the twelfth national conference on Artificial intelligence **1** (1994) 444–449
16. Wikipedia: Golden Balls: https://en.wikipedia.org/wiki/Golden_Balls
17. Rousseau, D., Sitkin, S., Burt, R., Camerer, C.: Not so different after all: A cross-discipline view of trust. Academy of Management Review **23**(3) (1998) 393–404
18. Sabater, J., Sierra, C.: Review on computational trust and reputation models. Artificial Intelligence Review **24**(1) (2005) 33–60
19. Castelfranchi, C., Falcone, R.: Trust Theory. 1 edn. John Wiley & Sons, Ltd, Chichester, UK (mar 2010)
20. Gambetta, D.: Can We Trust Trust? In: Trust: Making and Breaking Cooperative Relations. Blackwell (1988) 213–237
21. Marsh, S.P.: Formalising Trust as a Computational Concept. PhD thesis (apr 1994)
22. Castelfranchi, C., Falcone, R.: Principles of trust for MAS: cognitive anatomy, social importance, and quantification. Proceedings of the International Conference on Multi Agent Systems (1998) 72–79
23. Abdul-rahman, A., Hailes, S.: Supporting Trust in Virtual Communities. System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on **00**(c) (2000) 1–9
24. Sabater, J., Sierra, C.: Reputation and social network analysis in multi-agent systems. Proceedings of the first international joint conference on Autonomous agents and multiagent systems part 1 - AAMAS '02 (2002) 475
25. Sabater, J., Paolucci, M., Conte, R.: Repage: REPutation and ImAGE among limited autonomous partners. Jasss **9**(2) (2006) 117–134

26. Huynh, T.D., Jennings, N.R., Shadbolt, N.R.: An integrated trust and reputation model for open multi-agent systems. Autonomous Agents and Multi-Agent Systems **13**(2) (2006) 119–154
27. Pinyol, I.: Reputation-Based Decisions for Cognitive Agents (Thesis Abstract). Doctoral Mentoring Program (Aamas) (2009) 15–16
28. Nash, J.: Non-Cooperative Games. The Annals of Mathematics **54**(2) (sep 1951) 286
29. Han Yu, Zhiqi Shen, Leung, C., Chunyan Miao, Lesser, V.R.: A Survey of Multi-Agent Trust Management Systems. IEEE Access **1** (2013) 35–50
30. Pinyol, I., Sabater-Mir, J.: Computational trust and reputation models for open multi-agent systems: a review. Artificial Intelligence Review **40**(1) (jun 2013) 1–25
31. Noorian, Z., Ulieru, M.: The State of the Art in Trust and Reputation Systems: A Framework for Comparison. Journal of theoretical and applied electronic commerce research **5**(2) (aug 2010) 97–117
32. Huang, H., Zhu, G., Jin, S.: Revisiting Trust and Reputation in Multi-agent Systems. Computing, Communication, Control, and Management, 2008. CCCM '08. ISECS International Colloquium on **1** (2008) 424–429
33. Schaefer, K.: The Perception and Measurement of Human-Robot Trust. PhD thesis (2009)

Appendices

A Variants of Single-subject designs

B Landscape Appendice

landscape page