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 . . . 3
1.1 Goals . . . 3
2 Background . . . 3
2.1 Trust . . . 4
2.1.1 Castelfranchi and Falcone's Trust . . . 4
2.2 Reputation . . . 5
3 Related Work . . . 5
3.1 Trust Models . . . 5
3.1.1 Castelfranchi et. al. . . . 5
3.2 Discussion . . . 6
4 Proposed Solution . . . 6
4.1 Subsection Heading Here . . . 6
5 Evaluation . . . 7
6 Conclusion . . . 7

1 Introduction

Trust has been described in Psychology as being one of the most important components of interpersonal relationships [1]. So as Artificial Intelligence (AI) Research gravitates towards the development of Intelligent Agent Systems [2], out of which a focal concern is the performance of collaborative tasks [3–5]. and addressing the problems of interaction between humans and agents [6], Trust should be one of the main focus on Human Agent Interaction (HAI). 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 [7], specially when so much has been done in modelling Trust and Reputation in Multi-Agent System (MAS) [8], Trust in Automation [9–11] and in HRI [12,13].

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]. And to that goal we propose to develop two self-contained agent modules, capable of creating a cognitive model representing the mental state of the user's trust in the agent, and provide input on what actions should be used to improve or worsen trust on the agent. This will allow, not only to ensure proper reliance in the agents task [11], but to also provide additional insight on how certain actions affect human trust on agents differently from one another. It can also be used to improve Relational Agents by adding to the trust component of long-term relationships between Agents and Humans [16].

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 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 concerning Trust.

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 [10, 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 basis 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 [20]), 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 [21] 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 [22] 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 [21], 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 [23]. 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 [23] state that Trust can be defined with a central core, composed

by a five-part relation, between the trustor and 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 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 will purposely ruin the task and go against your goals.

2.2 Reputation

Reputation is also a concept that appears very often linked with trust in the literature, specially since most models created for representing trust have been focused on MASs, where trust is influenced not 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.

3 Related Work

Computational Trust research has been focused on modelling trust in MASs, specially on open e-commerce environments [8, 24–27], with at least 106 models created [8], since the formalization of trust as a measurable property by Marsh in 1994 [22].

3.1 Trust Models

For related work concerning Trust Models we will focus on Cognitive Trust Models, first introduced by Castelfranchi et. al. [23], 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,

The work I will be comparing my work with (direct influence, biggest part of this document, has to make sense, it has to show the relevance of mine work in a international way)

In data-driven timing generation for social behaviors, using Machine Learning (ML) techniques, it is retrieved two types of samples: positive samples representing the moments (or timings) in the interaction that are socially acceptable

to generate a backchannel gesture (e.g., a head nod or a vocalization "hmm hmm") and negative samples as the moments when it is unacceptable. In previous approaches, corpus based, positive samples are taken directly from the annotated dataset and the negative randomly as long as they do not overlap with a positive sample.

3.2 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 **??**.

According to previous studies [33 dont stare at me], during dyadic interactions, the listener usually maintains long gazes at the speaker and only interrupts briefly from time to time.

In fact, [9 toward dyadic...] found that in a negotiation setting not reciprocating negative self-disclosure led to decreased feelings of rapport. [**?**]

mutual gaze in determine turn-taking turn-taking [8] [12] [14] [56] [47] [todos do dont stare at me].

One of the most notable non-verbal behaviors to build rapport is gaze because it is a clear signal of mutual attention (as our parents said, look to people eyes when talking to them), acts as an invitation to interaction, increases dynamism, likelability and believability [4 do dont stare at me]

its impact in a wide range of interpersonal domains includ- ing social engagement [52], classroom learning [22], suc- cess in negotiations [20], improving worker compliance [18], psychotherapeutic effectiveness [59], and improved quality of child care [11].

Gaze as object of interest [8] [37] [55]. , effects on the way communication proceeds [54] [60] [23] [19] [28].

In fact, previous work demonstrated that there is evidence that in health domains, high rapport doctors engaged in less extensive eye-contact than low

rapport doctors , 85% and 70% of the interaction time respectively . However, the impact of the gaze depends if the interacts are in a helping context (e.g. meetings with a doctor) or in a non-helping context (e.g. interviewing) [dont stare at me]. On the latter, directed gaze is correlated positively with participant's evaluative impression [Tickle-Degnan and Rosenthal]. In interviweing ocntexts, Goldberg, Kiesler, and Collins [25] found that people who spent more time gazing at an interviewer received higher socio- emotional evaluations.

Argyle [1] found that in dyadic conversa- tions, the listener spent an average of about 75

Kendon [33] reported that a typical pattern of interaction when two people converse with each other consists of the listener maintaining fairly long gazes at the speaker, interrupted only by short glances away.

In short, gaze can also have negative impact if not dosed correctly.

To assess if negative arousal played some role, we asked the partici- pants to evaluate how uncomfortable they were when inte- racting with the agent (the embarrassment scale) in the post-questionnaire packet. ANOVA

5 Evaluation

Evaluation

6 Conclusion

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

Acknowledgment

The author would like to thank...

References

1. Simpson, J.A.: Psychological Foundations of Trust. Current Directions in Psychological Science **16**(5) (oct 2007) 264–268
2. Russell, S., Norvig, P.: Artificial Intelligence: A Modern Approach, 3rd edition. (2009)
3. Grosz, B.J.: Collaborative Systems. AI Magazine (1996) 67–85
4. Allen, J., Ferguson, G.: Human-machine collaborative planning. International NASA Workshop on Planning (2002) 1–10
5. Allen, J., Chambers, N., Ferguson, G., Galescu, L., Jung, H., Taysom, W.: PLOW : A Collaborative Task Learning Agent. Interpreting **22** (2007) 1514–1519
6. Bradshaw, J.M., Feltovich, P., Johnson, M.: Human-Agent Interaction. Handbook of HumanMachine Interaction (2011) 293–302
7. Bickmore, T.W., Picard, R.W.: Establishing and Maintaining Long-Term Human-Computer Relationships. ACM Transactions on Computer-Human **12**(2) (2005) 293–327

8. 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
9. Lee, J., Moray, N.: Trust, control strategies and allocation of function in human-machine systems. Ergonomics **35**(10) (oct 1992) 1243–70
10. Jones, S., Marsh, S.: Human-computer-human interaction. ACM SIGCHI Bulletin **29**(3) (jul 1997) 36–40
11. Lee, J.D., See, K.A., City, I.: Trust in Automation : Designing for Appropriate Reliance. **46**(1) (2004) 50–80
12. Goodrich, M.a., Schultz, A.C.: Human-Robot Interaction: A Survey. Foundations and Trends® in Human-Computer Interaction **1**(3) (2007) 203–275
13. 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)
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. Bickmore, T., Cassell, J.: Relational agents. In: Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '01. Volume PhD Thesis., New York, New York, USA, ACM Press (2001) 396–403
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. Simpson, J.a.: Foundations of interpersonal trust. In: Social psychology: Handbook of basic principles (2nd ed.). (2007) 587–607
21. Gambetta, D.: Can We Trust Trust? In: Trust: Making and Breaking Cooperative Relations. Blackwell (1988) 213–237
22. Marsh, S.P.: Formalising Trust as a Computational Concept. PhD thesis (apr 1994)
23. 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
24. 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
25. 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
26. 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
27. 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