The Impact of Explanation Dialogues on Human-Computer Trust

Florian Nothdurft, Tobias Heinroth, and Wolfgang Minker

Institute of Communications Engineering
University of Ulm, Germany
{florian.nothdurft,tobias.heinroth,wolfgang.minker}@uni-ulm.de

Abstract. Maintaining and enhancing the willingness of a user to interact with a technical system is crucial for human-computer interaction (HCI). Trust has shown to be an important factor influencing the frequency and kind of usage. In this paper we present our work on using explanations to maintain the trust relationship between human and computer. We conducted an experiment on how different goals of explanations influence the bases of human-computer trust. We present the results of the conducted study and outline what this means for the design of future technical systems and in particular for the central dialogue management component controlling the course and content of the HCI.

Keywords: Computer applications, Knowledge based systems, Cooperative systems, Adaptive systems, Expert Systems.

1 Introduction

Advances in human-computer interaction based technology enable the vision of mobile or ubiquitous technical systems accompanying users in their daily life. These systems have the possibilities to serve as a personal assistant due to the potential long-term relationship between a human and a technical system. Personal assistants should have the potential to solve complex problems the user is faced with daily or solely and which require significant interaction. However, this paradigm of interaction requires a working relationship between the human and the technical system. Such a relationship is characterized by a user's cooperativeness during interaction and his trust in the technical system.

Human-computer trust has shown to be a crucial point in keeping the user motivated and cooperative. The users' trust in a technical system will decrease if he does not understand system actions or instructions (Muir, 1992). This may lead to a change in the willingness to interact or in the worst case scenario to an abort in interaction and use (Parasuraman & Riley, 1997). However, providing explanations can help to prevent a decrease of trust (Glass, McGuinness, & Wolverton, 2008).

Similar to explanations in human-human interaction, explanations in human-computer interaction pursue a certain goal. Explanations are given to clarify, change or impart knowledge with the implicit idea to align and adapt the mental models of

M. Kurosu (Ed.): Human-Computer Interaction, Part III, HCII 2013, LNCS 8006, pp. 59–67, 2013.

the participating parties. In this case the users' constructed mental model of the system has to be adapted to the correct mental model (i.e. the designed behavior, knowledge and reasoning) of the system. This means, that not only the knowledge of the user has to be adapted to match the system required knowledge, but the behavior of the system has to be explained to the user in order to keep his transparency in the system. For example, if the users' believed reasoning process of the system does not match the real reasoning process, the occurrence of not understandable or as wrong perceived situations seems inevitable. The correction of the users' mental model can be done by providing explanations.

In order to pursue a particular objective an explanation goal (see Table 1 for a listing of explanation goals) has to be selected. However, a complex explanation can pursue more than one goal at a time. For example, one can think of a learning explanation which contains a conceptual explanation as well. Additionally a simple explanation does not necessarily have to pursue only one goal. The mapping of the presented final explanation to the inherited explanation goal does not have to be a one to one mapping. This means that the effect of a given explanation is not limited to the originally intended goal, but may implicitly pursue other goals as well.

Table 1. The different goals an explanation can pursue

Goal of Explanation	*Description*
Justification	Explain the motives of the answer?
Transparency	How was the systems answer reached?
Relevance	Why is the answer a relevant answer?
Conceptualization	Clarify the meaning of concepts
Learning	Learn something about the domain

For our experiment we concentrated on justification and transparency explanations. Justifications are the most obvious goal an explanation can pursue. The main idea of this goal is to provide support for and increase confidence in given system advices or actions. The goal of transparency is to increase the users understanding in how the system works and reasons. This can help the user to change his perception of the system from a black-box to a system the user can comprehend. By this, the user can build a mental model of the system and its underlying reasoning processes. Selecting the appropriate goal of explanation based on users' human-computer trust is an unprecedented approach because existing studies concentrate on trust as a one-dimensional concept. However, Trust is multi-dimensional and consists of several bases. For human relationships, Mayer (Mayer, Davis, & Schoorman, 1995) defined three levels that build the bases of trust: ability, integrity and benevolence.

For human-computer trust (HCT) Madsen and Gregor (Madsen & Gregor, 2000) constructed a hierarchical model (see Fig. 1). They tried to separate trust into nine basic constructs but eliminated four constructs because of representative or discriminative issues. This results in five basic constructs of trust, with two major components (cognitive- and affect-based components) and expected relationships modeled between them. However, as Mayer already stated, the bases of trust are separable, yet related to one another. All bases must be perceived high for the trustee to be deemed trustworthy. If any of the bases does not fulfill this requirement, the overall trustworthiness can suffer (Madsen & Gregor, 2000).

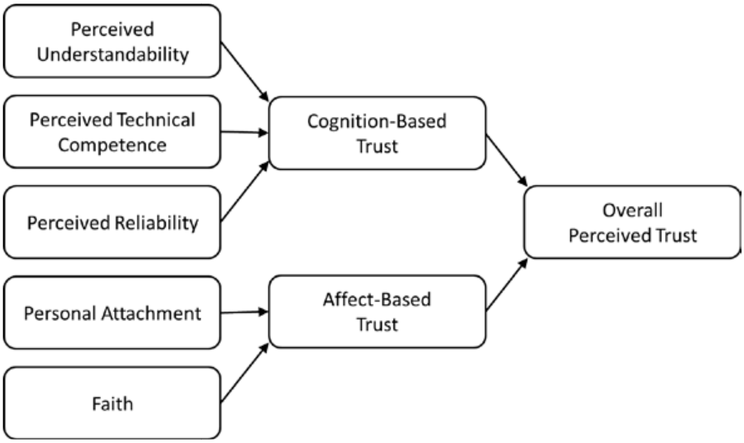


Fig. 1. In this for human-computer trust constructed model, personal attachment and faith build the bases for affect-based trust and perceived understandability, perceived technical competence and perceived reliability for cognition-based trust (Madsen & Gregor, 2000)

If we want to use explanations to influence the human-computer trust relationship in a directed and not arbitrary way, we need to find the most effective mapping of explanation goals to HCT bases (see Fig. 2). This means, that we have to identify which goal of explanation influences which base of trust in the most effective way. Thereby, undirected strategies to handle HCT issues can be changed into directed and well-founded ones, substantiating the choice and goal of explanation.

In our experiment we wanted to test how the different goals of explanation do influence the bases of trust in unexpected, not understandable situations in human-computer interaction (HCI). The main idea was to influence the HCT relationship in a negative way and to analyze how different explanation goals can help to remedy or reduce occurring trust issues to prevent the user from losing the willingness to interact with the system.

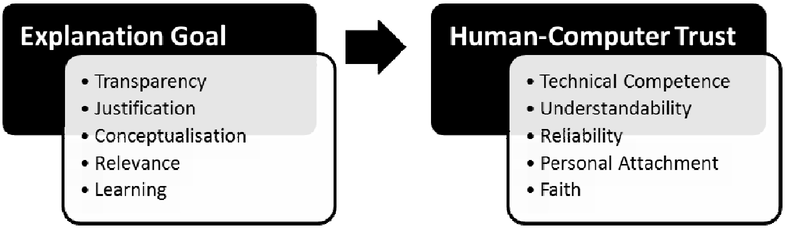
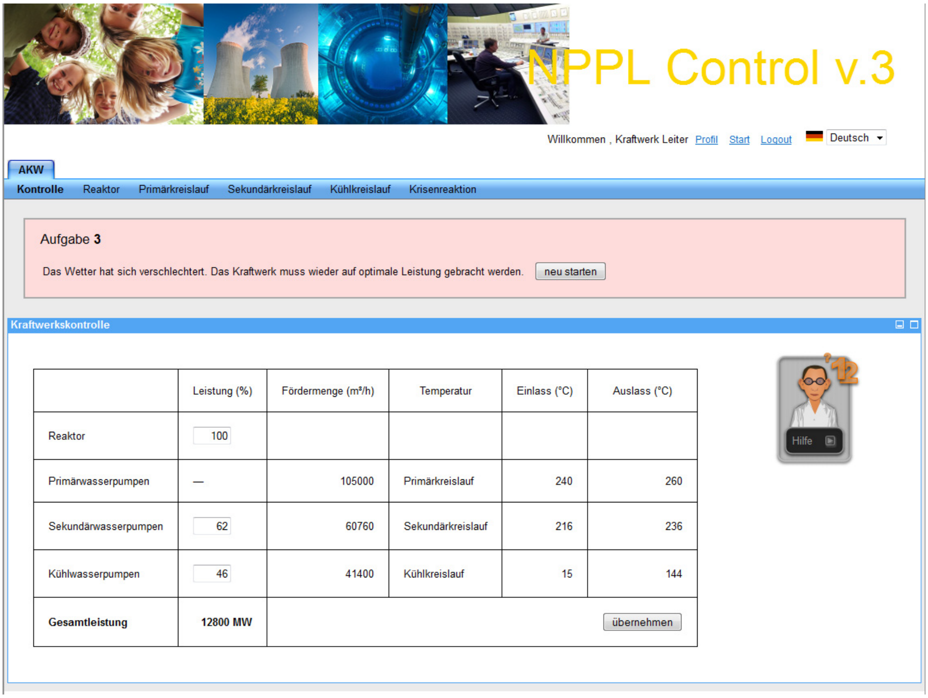


Fig. 2. The main goal is to find the most effective mapping between goals of explanation and the bases of human-computer trust

2 The Experiment

The setting of the experiment was a web-based simulation of a nuclear power plant control room. The subjects had to accomplish several rounds of interaction in which they had to solve pre-defined tasks. During those rounds they were assisted by a virtual anthropomorphic assistant (Lang & Minker, 2012) which helped the user proactively with the upcoming tasks. The user interface represented the controls of the nuclear power plant by distributing control room functionalities over various tabs (see Fig. 3).


NPPL Control v.3
Willkommen , Kraftwerk Leiter Profil Start Logout
Deutsch
AKW
Kontrolle Reaktor Primärkreislauf Sekundärkreislauf Kühlkreislauf Krisenreaktion
neu starten
Kraftwerkskontrolle
Hilfe
übernehmen

Aufgabe 3

Das Wetter hat sich verschlechtert. Das Kraftwerk muss wieder auf optimale Leistung gebracht werden.

	Leistung (%)	Fördermenge (m³/h)	Temperatur	Einlass (°C)	Auslass (°C)
Reaktor	100				
Primärwasserpumpen	—	105000	Primärkreislauf	240	260
Sekundärwasserpumpen	62	60760	Sekundärkreislauf	216	236
Kühlwasserpumpen	46	41400	Kühlkreislauf	15	144
Gesamtleistung	**12800 MW**				übernehmen

Fig. 3. This is a screenshot of the experiment interaction interface. On the right we can see the agent in sleep mode. Distributed over the tabs are the nuclear power plant functionalities. Beneath is the text describing the task the user has to accomplish. In this case the weather is getting worse and the power plant has to be adjusted to provide more power.

However, during selected rounds, the simulation reacted unexpectedly. The main idea behind this was to provoke a decrease in human-computer trust between the human and the machine. By this we wanted to test how different goals of explanation can help to prevent or reduce the expected trust loss. In total 60 test persons took part in the experiment. For each kind of explanation 20 persons were tested with an evenly distributed number of males and females. However, due to incomplete data only 48 valid subjects remained. The average age was 23.35 with the majority of the participants being students. In order to measure the influences on the bases of human-computer trust we were using a translated version of the working alliance inventory questionnaire modified by Madsen and Gregor (Madsen & Gregor, 2000) for the measurement of human-computer trust. The original questionnaire (Horvath & Greenberg, 1989) measures which trust and belief a therapist and patient have in each other in achieving a desired outcome. The HCT questionnaire was adapted to our needs and consisted of 15 items (three items for each base of trust).

In total the subjects had to complete seven tasks. For example, the nuclear power plant had to be controlled to output a certain amount of power. However, three of the seven tasks were interrupted by unexpected situations. For example, a water pump was broken or some control rod elements were defect. These situations were meant to be incongruent to the users' mental model of the system and therefore not understandable and unexpected. The course of the experimental design can be seen in Table 2.

Table 2. The Course of the experiment regarding unexpected situations and provided HCT-questionnaires

Round	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Error / unexpected Situation			x		x	x	
HCT Questionnaire		x	x	x	x	x	

In the beginning we wanted the user to accustom to the system. Therefore, the first questionnaire was presented after the second task. In the third, fifth and sixth round the task was interrupted by an unexpected system error (see Fig. 3).

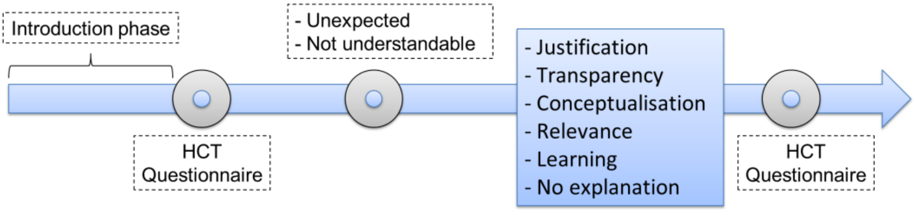


Fig. 4. Rounds with unexpected situations in the experiment: In the first study we examined justification, transparency and no explanation

As mentioned before the system reaction was either augmented by a transparency or justification explanation. The baseline was a group provided with no additional explanation. These experiments allowed us to determine, whether our constructed unexpected situations did influence the HCT negatively. Our hypothesis was that both goals of explanation would perform better in terms of keeping trust than no explanation at all. Additionally, we assumed that justification explanations would help especially the bases of technical competence and understandability. For transparency explanations we expected influences on the bases of understandability and reliability.

3 Results

The first problem we encountered was that the unexpected situations did not induce the anticipated trust loss. In our opinion, this was either due to the too good interaction and help of the system represented by the virtual agent or the introductory phase was too short to build a trust relationship between man and machine.

Observing the data we did not find any significant differences between providing the user with no explanations, justifications or transparency explanations. Especially the development over time seemed rather arbitrary in terms of system errors influencing the human-computer trust relationship. However, when analyzing the data we found some gender differences (see Fig. 4).

Concerning the base of the perceived reliability we found a marginal significant ($H^2 = 2.9, p < .08$) difference when using transparency explanations (4.29 for males to 3.49 for females). When providing justifications we got a significant difference ($H^2 = 4.0, p < .05$) concerning the perceived faith between males (3.2) and females (4.12). When observing only female subjects we could prove a marginal significant ($U = 8.5, p < .09$) difference between providing no explanation (3.9) compared to providing justifications (4.57) for the base of reliability.

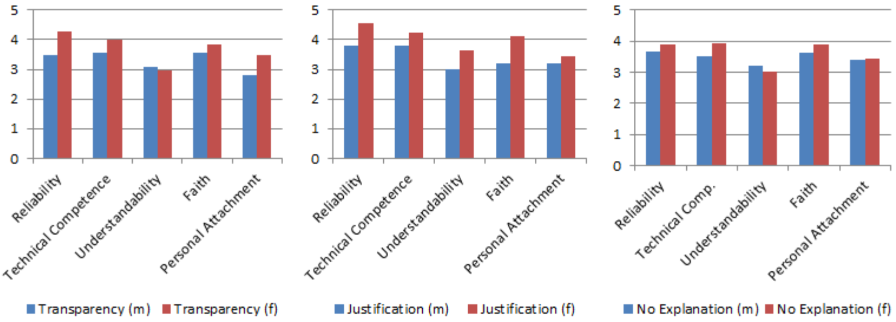


Fig. 5. Results for the different types of intervention divided on gender

Closer examination of the data revealed some further tendencies. However, due to the study design there were not sufficient numbers of subjects to analyze the male to female differences to draw valid conclusions. In the next chapter we will discuss the results and mention some of the rather interesting tendencies which we hope to address in more detail in future work and experiments.

As mentioned before the situations meant to influence HCT negatively did not serve their purpose. Despite experiencing not understandable situations the help provided by the agent was sufficient to handle the occurring problem. The agent provided a step-by-step tutorial on how to overcome the experienced problems. Therefore, in a follow up experiment we want to separate the occurring system error from the task the subject has to accomplish. Additionally, we plan to extend the introductory phase. Every type of task the user has to execute later on, should be done in a comparable way in the beginning. This way we hope to build a more complete users' mental model of the system.

Taking a closer look at the development of perceived reliability and perceived technical competence (see Fig. 5) when providing transparency explanation, we observe that for male subjects the curve has rather negative tendencies. Compared to that, females seem to be influenced only in a marginal way by transparency explanations, at least in our experiment.

Providing justifications seemed to benefit the perceived reliability regardless of gender. For perceived technical competence females seemed to benefit from justifications (as well see Fig. 5). As we lost in total 10 female subjects to incomplete questionnaires or quitting of the experiment (compared to 2 males), the number of females per explanation goal was limited. However, the results give some evidence that the gender aspect of explanation goals is worth more thorough investigation in the context of our follow-up studies.

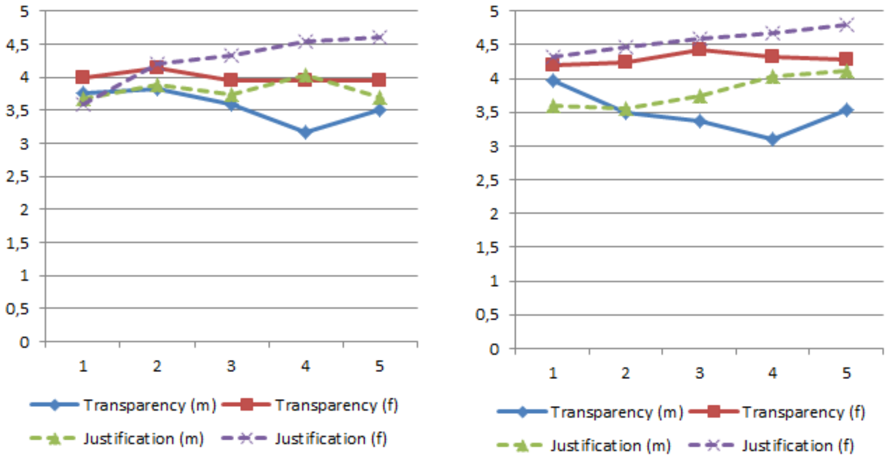


Fig. 6. The values of perceived technical competence on the left and perceived reliability on the right as given by the HCT questionnaire

Despite the setback of not influencing the HCT negatively, we can state that different goals of explanation do influence the bases of trust in a particular way. For example, justifications do influence the perceived reliability of the user towards the system (i.e. Fig. 5). Despite being not significant, this could be a first indicator that goals of explanation influence particular bases of trust.

4 Discussion

Technical systems meant to accompany and assist the user in his daily life are prone to the consequences of users' loss of trust. Personal assistants rely on users' cooperativeness and motivation to interact with the system and share relevant information. This is necessary to facilitate the adaptation and personalization of the technical system to the individual user. Especially future technical systems will have to adapt to a user's capabilities, preferences, requirements, and current needs. As these continually available assistants should appear competent, co-operative, and reliable to their users the relationship between human and technical system has to be healthy and trustworthy. If technical systems should change from helpers for simple, domain dependent tasks to individualized continually available assistants for complex and wide-spread domain independent tasks the dialogue between human and technical systems can be a typical bottleneck. The accomplishment of complex tasks requiring extensive dialogue depends on the users' willingness and motivation to interact. Maintaining the trustworthiness of the systems is, as mentioned earlier, one of the key factors to keep the user willing to interact with the system. Continually available and assisting systems possess a bunch of relevant information on the user, which can facilitate the preservation of human-computer trust. Gender, education, technical background and history of interaction can help to recognize situations prone to HCT-decrease and to select the probably most suitable intervention to remedy possible negative effects.

As presented in this paper, we found significant results as well as important evidence that the loss of trust has to be handled in a directed, not arbitrary way. Diverse types of explanations can influence the bases of trust in directed way. A well-founded selection of the most effective type of explanation will help to handle occurring trust issues in a non-arbitrary way. Additionally, the experiment indicates that females and males seem to react in a different way to types of provided explanations. As this does not have to be the only additional factor in the explanation selection, future research has to take other factors as education or technical background into account as well. Especially technical background is known to have an effect on explanations. The content, complexity and form of the explanation have to be adapted to the level of technical background and knowledge a user possesses. For example, expert and novice users profit differently from the diverse goals of explanation (Lim, Dey, & Avrahami, 2009). In our opinion, the goals of explanation may influence the bases of trust differently regarding the expertise as well.

However, there are still some missing pieces to be able to use these and upcoming results in a technical system for adaptation and individualization. Knowing how to react in the probably most effective way to a decrease in a base of trust is important, but the information which base of trust was influenced is essential as well. For this we need to categorize the context information in situations of trust loss. If we can state with a certain probability, that a particular context along with unexpected situations influences a specific base or bases of trust, we can react in an appropriate manner. This will enable us to include the results in an existing architecture to handle human-computer trust issues by providing explanations (Nothdurft, Bertrand, Lang, & Minker, 2012).

5 Conclusion and Future Work

In this paper we outlined why it is necessary to consider human-computer trust in technical systems. Especially in systems meant to be personal, individualized and continually available assistants, the long-term relationship between human and technical system is highly relevant. Unexpected or not understandable situations have shown to influence the users' HCT negatively. Research has shown that explanations can help to remedy negative effects occurring in these situations. However, the mapping between the different goals of explanations and the diverse components of human-computer trust remains unclear. Therefore, we presented an experiment on how different goals of explanation influence particular bases of human-computer trust. We found indication that indeed differences exist in the effects of explanation goals on the bases of trust. Additionally, we did find some gender aspects, which seem to be worth analyzing more extensively in follow-up experiments. Therefore, it seems worthy to consider individual and situational parameters when providing explanations in critical situations.

Acknowledgements. This work was supported by the Transregional Collaborative Research Centre SFB/TRR-62 “Companion-Technology for Cognitive Technical Systems” which is funded by the German Research Foundation (DFG). Additional thanks to M. Attenberger, L. Knecht, N. Maucher, M. Partsch and F. Rueck for conducting the study.

References

Glass, A., McGuinness, D.L., Wolverton, M.: Toward establishing trust in adaptive agents. In: IUI, pp. 227–236. ACM (2008)

Horvath, A., Greenberg, L.: Development and validation of the Working Alliance Inventory. Journal of Counseling Psychology, 223–233 (1989)

Jörg, C., Anders, K.-P.: Designing Explanation Aware Systems: The Quest for Explanation Patterns. In: Explanation-Aware Computing – Papers from the 2007 AAAI Workshop, pp. 20–27. AAAI Press, Vancouver (2007)

Lang, H., Minker, W.: A collaborative web-based help-system. In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics, pp. 61–65. ACM, Craiova (2012)

Lim, B.Y., Dey, A.K., Avrahami, D.: Why and why not explanations improve the intelligibility of context-aware intelligent systems. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 2119–2128. ACM, Boston (2009)

Madsen, M., Gregor, S.: Measuring human-computer trust. In: 11th Australasian Conference on Information Systems (2000)

Mayer, R.C., Davis, J.H., Schoorman, F.D.: An Integrative Model of Organizational Trust. The Academy of Management Review, 709–734 (1995)

Muir, B.M.: Trust in automation: Part I. Theoretical issues in the study of trust. In: Ergonomics, pp. 1905–1922 (1992)

Nothdurft, F., Bertrand, G., Lang, H., Minker, W.: Adaptive Explanation Architecture for Maintaining Human-Computer Trust. In: 36th Annual IEEE Computer Software and Applications Conference (COMPSAC 2012), pp. 176–184. IEEE, Izmir (2012)

Parasuraman, R., Riley, V.: Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors. The Journal of the Human Factors and Ergonomics Society, 230–253 (1997)

Sormo, F., Cassens, J.: Explanation goals in case-based reasoning. In: Proceedings of the ECCBR 2004 Workshops, pp. 165–174 (2004)