Bases of Human-Computer Trust and Explanations

Florian Nothdurft
Institute of Communications Engineering
Ulm University, Germany
florian.nothdurft@uni-ulm.de

Helmut Lang
Institute of Communications Engineering
Ulm University, Germany
helmut.lang@uni-ulm.de

Melina Klepsch
Institute of Psychology and Education
Ulm University, Germany
melina.klepsch@uni-ulm.de

Wolfgang Minker
Institute of Communications Engineering
Ulm University, Germany
wolfgang.minker@uni-ulm.de

CHI 2013 Extended Abstracts, April 27–May 2, 2013, Paris, France.
ACM 978-1-4503-1952-2/13/04.

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 ongoing work on using explanations to maintain the trust relationship between human and computer. We describe an experiment on how different goals of explanations can be used to influence the bases of human-computer trust in a directed way. We present the results of a conducted preliminary study and outline how to improve the experiment so as to be able to include the results in an existing dialogue system.

Author Keywords
Human-Computer Trust; Explanations; Dialogue Systems; Adaptive Systems; Expert Systems

ACM Classification Keywords
H.5.2 [Information interfaces and presentation (e.g., HCI)]: User-centered design

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 users 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.

One important information during interaction between human and computer is the trust model of the user (see fig. 1). Trust has shown to be a crucial point in keeping the user motivated and cooperative. The users' trust in a technical system will be decreased if he does not understand system actions or instructions [6]. This may lead to a change in the willingness to interact or in the worst case scenario to an abort in interaction and use [8]. However, providing explanations can help to prevent a decrease of trust [1].

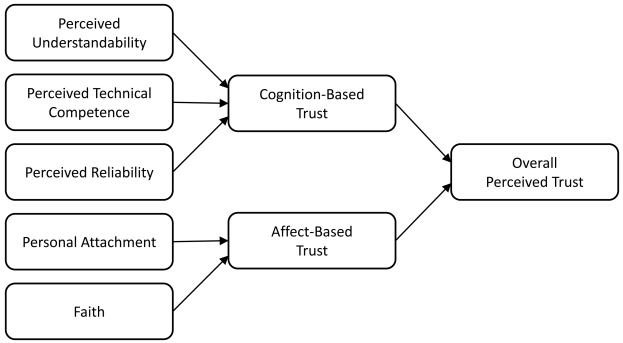

Figure 1: Human-computer trust model: Personal attachment and faith build the bases for affect-based trust and perceived understandability, perceived technical competence and perceived reliability for cognition-based trust.

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 the mental models of the participating parties. In order to pursue a particular objective an explanation goal (see table 1 for a listing of explanation goals) has to be selected.

Goals	Details
Transparency	How was the systems answer reached?
Justification	Explain the motives of the answer?
Relevance	Why is the answer a relevant answer?
Conceptualisation	Clarify the meaning of concepts
Learning	Learn something about the domain

Table 1: Goals of explanation after [9].

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 [5] defined three levels that build the bases of trust: ability, integrity and benevolence.

For human-computer trust (HCT) Madsen and Gregor [4] constructed a hierarchical model (see figure 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 [4].

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. 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.

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 [3] 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 figure 2).

Explanation	Males	Females
Transparency	10	7
Justification	9	5
None	9	8

Table 2: The distribution of test persons among the kinds of explanation and gender after data clearing.

Task	Error	Questionnaire
1	-	-
2	-	yes
3	yes	yes
4	-	yes
5	yes	yes
6	yes	yes
7	-	-

Table 3: The course of the study.

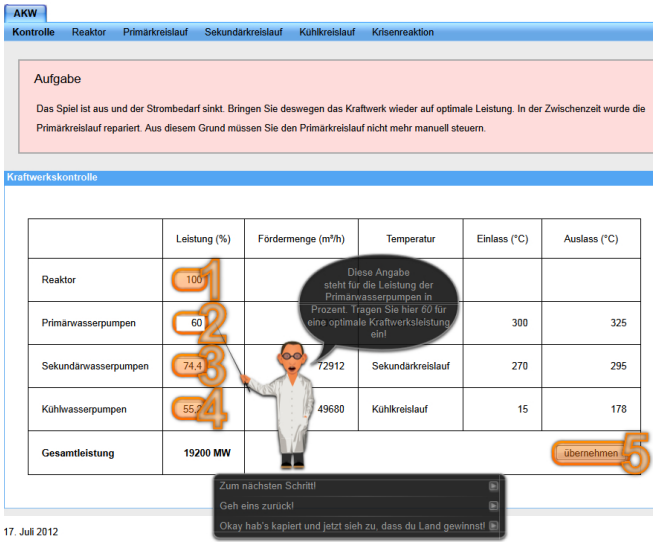

Figure 2: On top are the various tabs representing the power plant control room with a box beneath presenting the current task the user has to accomplish. The agent points on controls and provides help for the user.

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

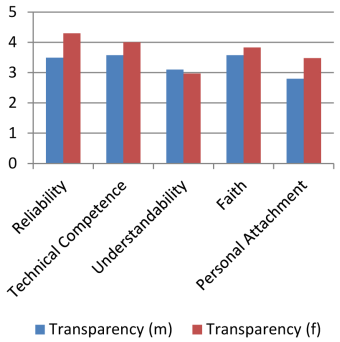

Figure 3: HCT for transparency explanations.

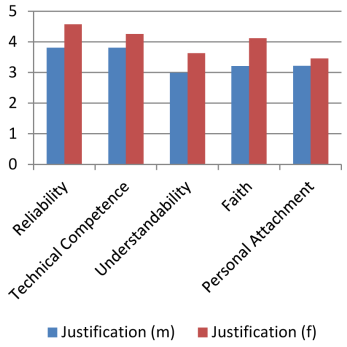

Figure 4: HCT for justifications.

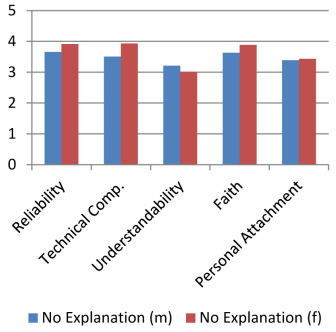

Figure 5: Baseline HCT.

evenly distributed number of males and females. However, due to incomplete data only 48 valid subjects remained (see table 2 for details). 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 modified "working alliance inventory" questionnaire by Madsen and Gregor [4]. The original questionnaire [2] measures which trust and belief a therapist and patient have in each other in achieving a desired outcome. This 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 3.

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. 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.

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. Concerning the base of the perceived reliability we found a marginal significant ($H(2)= 2.9$, $p_1<.08$) difference when using transparency explanations (4.29 for males to 3.49 for females).

When providing justifications we got a significant difference ($H(2)= 4.0$, $p_1<.05$) concerning the perceived faith between males(3.2) and females(4.12).

When observing only female subjects we could prove a marginal significant ($U= 8.5$, $p_1<.09$) difference between providing no explanation (3.9) compared to providing justifications (4.57) for the base of reliability.

Closer examination of the data revealed some further tendencies. However, due to the study design there were

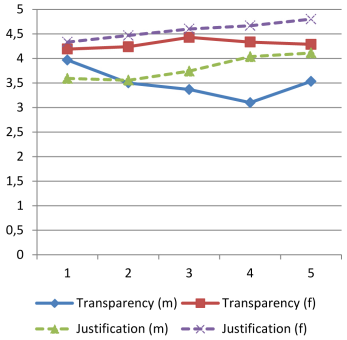

Figure 6: Perceived Reliability.

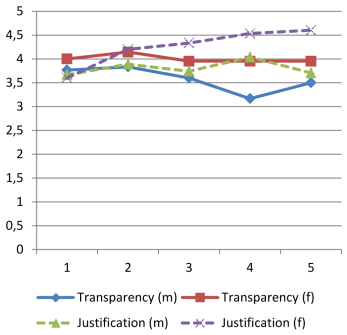

Figure 7: Perceived Technical Competence.

not sufficient numbers of subjects to analyze the male to female differences allowing 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.

Explanation	Reliab.	Tech.C.	Underst.	Faith	Pers.A.
Transparency (m)	3.49	3.57	3.1	3.57	2.79
Transparency (f)	4.29	4	2.97	3.83	3.48
Justification (m)	3.8	3.81	2.99	3.2	3.21
Justification (f)	4.57	4.25	3.62	4.12	3.45
No Explanation (m)	3.65	3.5	3.21	3.62	3.38
No Explanation (f)	3.9	3.93	3.01	3.88	3.43

Table 4: The average results of HCT-Bases over the course of five questionnaires.

Discussion

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 (see figure 6) and perceived technical competence (figure 7) 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 (see figure 6). For perceived technical competence females seemed to benefit from justifications (figure 7). 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.

Despite the set back 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. figure 6). Despite being not significant, this could be a first indicator that goals of explanation influence particular bases of trust.

However, as we were not able to influence the trust negatively in general, of course we could not influence the bases separately in a negative way. Therefore, we will use the present study as preliminary experiment. In the follow-up experiment, we want to influence the bases of trust in a directed way by considering the context of interaction in a more extensive way. 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 would enable us to include the results in an existing architecture to handle human-computer trust issues by providing explanations [7].

Conclusion and Future Work

In this paper 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, that seem to be worth analyzing more extensively in follow-up experiments. Therefore, we will use the experiment as preliminary study and its results to improve the follow-up experiment. The present study helped us a lot to focus on crucial points in designing experiments for human-computer trust and we hope that other researchers can benefit from it, too.

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

[1] A. Glass, D. L. McGuinness, and M. Wolverton. Toward establishing trust in adaptive agents. In *IUI '08: Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236, New York, NY, USA, 2008. ACM.

[2] A. O. Horvath and L. S. Greenberg. Development and validation of the working alliance inventory. *Journal of Counseling Psychology*, 36(2):223–233, 1989.

[3] H. Lang and W. Minker. A collaborative web-based help-system. In *Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, pages 60:1–60:5, New York, NY, USA, 2012. ACM.

[4] M. Madsen and S. Gregor. Measuring human-computer trust. In *Proceedings of the 11 th Australasian Conference on Information Systems*, pages 6–8, 2000.

[5] R. C. Mayer, J. H. Davis, and F. D. Schoorman. An Integrative Model of Organizational Trust. *The Academy of Management Review*, 20(3):709–734, 1995.

[6] B. M. Muir. Trust in automation: Part i. theoretical issues in the study of trust and human intervention in automated systems. In *Ergonomics*, pages 1905–1922, 1992.

[7] F. Nothdurft, G. Bertrand, H. Lang, and W. Minker. Adaptive explanation architecture for maintaining human-computer trust. In *36th Annual IEEE Computer Software and Applications Conference*, COMPSAC, June 2012.

[8] R. Parasuraman and V. Riley. Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 39(2):230–253, June 1997.

[9] F. Sørmo and J. Cassens. Explanation goals in case-based reasoning. In *Proceedings of the ECCBR 2004 Workshops*, 2004.