User Trust in Intelligent Systems: A Journey Over Time

Daniel Holliday
City University London
Centre for HCI Design
London, United Kingdom
daniel.holliday.1@city.ac.uk

Stephanie Wilson
City University London
Centre for HCI Design
London, United Kingdom
s.m.wilson@city.ac.uk

Simone Stumpf
City University London
Centre for HCI Design
London, United Kingdom
simone.stumpf.1@city.ac.uk

ABSTRACT
Trust is a significant factor in user adoption of new systems. However, although trust is a dynamic attitude of the user towards the system and changes over time, trust in intelligent systems is typically captured as a single quantitative measure at the conclusion of a task. This paper challenges this approach.

We report a case study that employed a combination of repeated quantitative and qualitative measures to examine how trust in an intelligent system evolved over time and whether this varied depending on whether the system offered explanations. We discovered different patterns in participants' trust journeys. When provided with explanations, participants' trust levels initially increased, before returning to their original level. Without explanations, participants' trust reduced over time. The qualitative data showed that perceived system ability was more important in determining trust amongst with-explanation participants and perceived transparency was a greater influence on the trust of participants who did not receive explanations. The findings provide a deeper understanding of the development of user trust in intelligent systems and indicate the value of the approach adopted.

Author Keywords
Trust; Intelligent Systems; Qualitative Evaluations; Quantitative Evaluations; Case Study

ACM Classification Keywords
H.1.2 User/Machine Systems: Human Factors

INTRODUCTION

The Need for Trust in Intelligent Systems
While intelligent systems are designed to aid users, they are not without their problems. Intelligent systems are dynamic and, through the use of machine learning algorithms, seek to adapt to the need and preferences of individual users. In doing so, they will typically violate these fundamental usability principles [6]: 1) control – an intelligent system may modify its behavior without explicit authorization from the user; 2) predictability – an intelligent system may produce a different output, when given the same input, at different points in time; 3) transparency – an intelligent system may not provide any understanding of its inner workings and so its behavior is not comprehensible to the user.

System ability, control, predictability and transparency have been identified in existing research as some of the key factors that influence users' trust in intelligent systems [2, 5, 12, 14]. Trust is a significant factor in determining the adoption of systems that perform tasks on behalf of the user [11]. Research has placed too great an emphasis on the components and machine learning algorithms of intelligent systems, with insufficient consideration for how users trust their actions [4]. Therefore, failure to adequately consider and address these factors in the design of an intelligent system is likely to result in a system that is not adopted by users.

What is Trust?
A vast number of definitions of trust exist. It has been studied from various perspectives across a range of disciplines and there is little consensus as to what trust means [10]. Trust is a "highly complex and multi-dimensional phenomenon," in which it is insufficient simply to ask an individual whether they trust or distrust another agent, as they may trust them in some regards but not in others [9]. Trust needs to be considered holistically and examined in its constituent parts rather than simply as a whole [3]. While an individual's initial judgments regarding the trustworthiness of another are founded in their general disposition to trust, their trust changes in response to the degree to which subsequent interactions either confirm or discredit those judgments [7]. Trust beliefs may grow or change over time with repeated interactions [13].

Existing Measures of Trust in Intelligent Systems
While existing research examines the multifaceted nature of user trust in intelligent systems, it does not consider the effect of repeated interactions on trust. In the context of intelligent systems, trust, and its associated factors, is

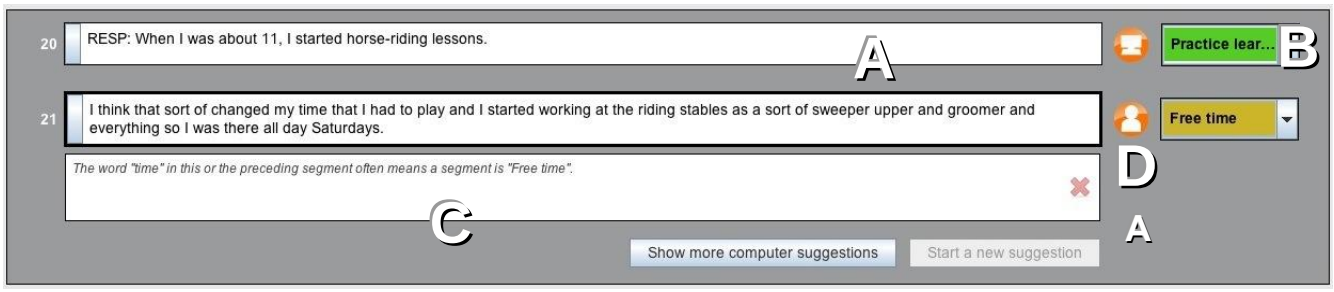

Figure 1. The AutoCoder user interface showing (A) a series of segments, (B) their corresponding codes, (C) a system-generated explanation, (D) an indication of whether the code has been assigned to a segment by the user or by the system.

measured solely in post-task questionnaires and interviews [2, 12, 14]. This current approach captures users' trust in intelligent systems only at a specific point in time. They give no representation of how trust and the factors of trust may have developed and changed over time. While users' trust journeys have received some consideration in other domains [15], intelligent systems differ from other technologies in a crucial aspect. The behavior of intelligent systems may change over time as the learning agent receives more training data, which, in turn, will affect user trust.

This paper proposes a combined iterative quantitative and qualitative approach to measuring user trust, and factors associated with trust, over time in intelligent systems. The effect of explanations on user trust has been examined using existing methods [1, 12, 16]. It follows to use explanations, a mechanism with which to potentially influence trust, to explore new approaches of measuring the development of user trust over time in intelligent systems.

METHOD

We set out to investigate users' trust journeys both in cases where the intelligent system offered explanations and in cases where it did not do so. A case study was devised and carried out in which participants were exposed to an intelligent system. 15 participants (7 male, 8 female), consisting of full-time postgraduate students and non-academic university staff across a range of different university departments, were assigned to one of two condition groups: 1) with-explanations; 2) without-explanations. From hereon, participants in the with-explanation and without-explanation condition groups will be referred to with participant codes beginning with 1 and 2 respectively.

The Intelligent System Used in the Case Study

The intelligent system used in the case study was the AutoCoder [8]. The AutoCoder is an intelligent assistant designed to aid users in the task of coding qualitative data, i.e. the process of marking segments of a transcript with codes (descriptive words or category names) for analysis in qualitative research. Figure 1 shows the graphical user interface of the AutoCoder and highlights its key features. The AutoCoder automatically codes the segments of the transcript. If the user disagrees with the code allocated by the system, they may correct it to the code that they judge appropriate. The AutoCoder produces system-generated explanations to explain the reasoning behind automatically coded segments. These explanations are based on words, combinations of words and punctuation. Participants in the with-explanation condition group used the AutoCoder as described above. Participants assigned to the without-explanation condition used a modified version of the system in which they did not receive explanations.

Procedure

Participants were asked to code an interview transcript using the AutoCoder. They were asked to think aloud as they carried out the coding task. After participants had manually coded three segments of the transcript, the AutoCoder displayed an alert instructing them complete an trust assessment (TA). Participants were asked to indicate the extent to which agreed with the statement, "I trust the AutoCoder to assist me in coding the remaining segments of the transcript." Participants expressed their agreement on a 7-point Likert scale, ranging from 1 (strongly disagree) to 7 (strongly agree). After participants had completed the TA and dismissed the alert, the AutoCoder automatically coded the remaining transcript segments. Participants then worked their way through the rest of the transcript, correcting or agreeing with the intelligent system's decisions. The TA was triggered at regular intervals during the remainder of the coding task; capturing participants' trust levels at these times. Following the coding task, participants completed a post-task questionnaire (PTQ), where again, participants' trust levels were recorded.

Quantitative and qualitative measures were obtained throughout the duration of the task. Participants' trust levels were captured at regular intervals through questions administered at each of the TAs and the PTQ. Their attitudes towards overall trust and the identified factors of trust (perceived system ability, perceived control, perceived predictability and perceived transparency) were gathered by asking them to think aloud during the coding task. Prior to the task, participants were instructed to describe what they were doing, why they were doing it, what they thought the system was doing, why they thought the system was doing what it was doing and any other thoughts that occurred to

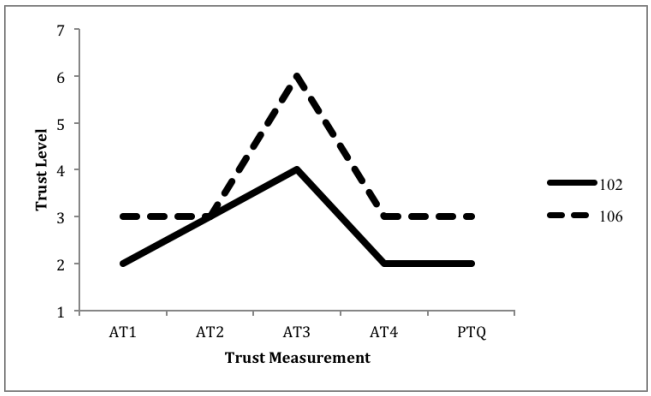

Figure 2. Graph showing the trust journeys of participants 102 and 106 of the with-explanation condition group.

them. During the session, participants received no prompts beyond efforts to encourage them to continue thinking aloud, such as, "okay," and, "please continue".

In the following section, we will describe a subsection of our data that exemplifies and emphasizes the worth of this combined quantitative and qualitative iterative approach to measuring trust over time in intelligent systems.

RESULTS AND DISCUSSION

The iterative measurement of trust at intervals throughout the duration of the coding task brought to light the different patterns in participants' trust journeys that were exclusive to each of the condition groups. Participant attitudes gleaned from the think aloud help us to understand the reasons for these variations. Participants took between approximately 60 – 90 minutes to complete the task.

We now consider the development of trust throughout the duration of the task. Two participants have been chosen from each condition group and their trust journeys discussed in detail, illuminated by the qualitative data. These participants' trust journeys were representative of each condition group. Participants 102 and 106 from the with-explanation condition and participants 204 and 206 from the without-explanation condition were selected as their think aloud qualitative data was the most illuminating, giving useful insight into their trust journeys. We would not be able to discuss the trust journeys of all participants in detail and without negating the value of the rich qualitative data in the scope of this paper. In examining the responses and attitudes of these participants in detail, we identify the factors of trust that most greatly influence their overall trust.

With-Explanation Condition Group Pattern – "Hump"

Figure 2 shows the similar trust journeys followed by participants 102 and 106 of the with-explanations condition group. The participants' initial trust in the AutoCoder was low, before increasing to reach a peak in the middle of the task. Their trust then decreased, returning to its initial level.

At TA1, the AutoCoder had not automatically coded any segments of the transcript, and so, this indicates participants' propensity to trust in the intelligent system. Participant 102 was skeptical of its abilities without evidence to the contrary: *"I wouldn't trust it until I see results. Instinctively, I wouldn't trust it."* Participant 106 was somewhat less hesitant to trust the system, stating he had, *"no reason to trust it but I'm quite confident for some reason. I wouldn't have a reason to distrust it, as such, either."*

While participant 102 perceived the system's ability to have improved, her trust did not yet increase, as she believed more time was needed for the AutoCoder to learn from more training data: *"[…] it is a learning system and it has learned better how to better answer […] I think it needs a bit more time to learn."* Participant 106 expressed similar sentiments regarding perceived system ability, although this was sufficient for his trust to increase: *"It is getting better but […] I think it's got to learn a little bit more."*

At TA3, the trust levels of both participants 102 and 106 reached their highest point. Participant 102 believed that the system had learned to produce a more accurate output based on her own input: *"[…] it seems like it's learned better,"* and participant 106 deemed that the system's ability had improved considerably and that it was, *"really, really getting there."*

However, both participants' trust decreased at TA4. Participant 106 believed that his trust was misplaced and asserted the AutoCoder's ability had deteriorated: *"I think I got a bit optimistic for the AutoCoder […] it all started to go a little bit wrong […] so that's knocked my confidence in it."* Similarly, participant 102's trust was influenced by their perceived ability of the system: *"I wouldn't trust it because […] I'm looking at how it's answered."*

Without-Explanation Condition Group Pattern – "Downward"

Figure 3 shows the similar trust journeys followed by participants 204 and 206 of the without-explanations

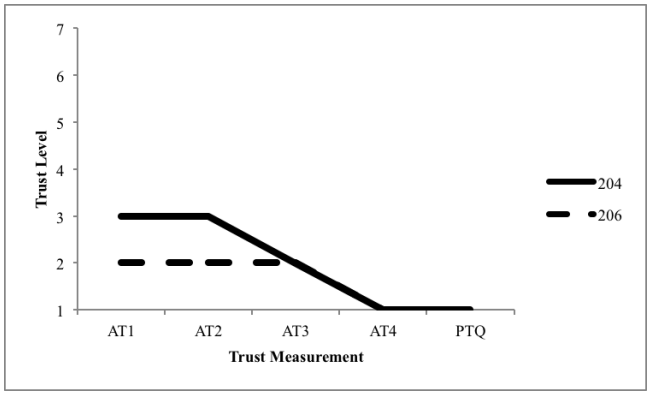

Figure 3. Graph showing the trust journeys of participants 204 and 206 of the without-explanation condition group.

condition group. The participants' initial trust in the intelligent system is low but this is, in fact, their highest level of trust in the entire task. This decreases, with both following the same decline in trust until the task's conclusion.

The participants' propensity to trust in the AutoCoder was captured at TA1, before it had automatically coded the transcript segments. Participant 204's initial trust level was influenced by the belief that the perceived system ability would be low: *"it will not have had the opportunity yet for enough learning."* Participant 206 was concerned that without having seen the AutoCoder in action, he would be unable to determine how it made its coding decisions: *"I don't know how it works […] I'm not sure what's going on in the background, you know."*

At TA2, both participants expressed frustration with their inability to understand the AutoCoder's workings due to the poor perceived transparency of the system, with participant 206 stating, *"I just don't understand why it's making the decisions it's making […] I think that's my biggest problem with it."* Participant 204 declared, *"I don't know at the moment how it's making its decisions."*

The participants frustration at poor perceived transparency in the intelligent system continued at TA3. Participant 206's trust remained constant, although he was unable to determine the workings of the AutoCoder: *"I don't know how it's doing it. Unless there's something else I'm not seeing, something deeper."* Participant 204's trust reduced when he, also, was still unable to determine how the system was making its decisions: *"I think there is a pattern, probably, but I haven't properly picked it up."*

The trust levels of both participants followed the same downward direction at TA4. Participant 204 attributed this to continued poor perceived transparency of the AutoCoder: *"I feel as though I ought to have picked up on more of a pattern […] about what I put in and what the machine does […] I really don't know."* Participant 206, also, was unable to determine the reasoning of the intelligent system, *"'cause recently it seems to have thrown a few random ones in there."*

The iterative measurement of participants' trust levels over time revealed different trust journeys that were exclusive to each of the condition groups. Through examination of the think aloud, the decrease in trust amongst participants in the with-explanation condition can be explained by a reduction in perceived system ability, that is, the perception that the AutoCoder's output was inaccurate. Meanwhile, the reduction in trust of participants in the without-explanations condition is founded in poor perceived transparency, that is, that the rationale behind the decisions made by the AutoCoder could not be understood.

While, in this instance, explanations did not result in a statistically significant difference in overall trust, we can observe differences in the factors of trust and their influence on overall trust. Indeed, emphasizing the contextual importance of perceived transparency over perceived system ability, participant 206 asserted, *"I know it's not going to be perfect and get everything right all the time but I don't know how it's doing it."*

CONCLUSION

The trust levels of participants in the with-explanation condition group were influenced strongly by the perceived ability of the AutoCoder. The presence of explanations assisted participants in the conception of an accurate mental model, increasing perceived transparency and the high-level understanding that the AutoCoder is a *"learning system."* Participants' trust in the AutoCoder in the without-explanation condition group was affected by the perceived transparency of the system. Without the presence of explanations, participants found it more difficult to conceive a mental model as to how the AutoCoder worked and, as a result, their level of trust in it did not increase. The think aloud data allow us to understand the 'why's' rather than just the 'what's' of individual users with regards to their trust in an intelligent system.

The findings of this case study exemplify the deeper, richer understanding of the development of users' trust over time in intelligent systems that can be obtained through the application of this combined iterative quantitative and qualitative approach. Using the single PTQ quantitative measure of trust at the conclusion of the task, it would have been easy to conclude, erroneously, that explanations had no effect on user trust or the factors of trust. The approach employed in this paper reveals how and why users' trust evolves over time, rather than simply what their level of trust is at a single point in time.

The trust level patterns of participants found exclusively in the with-explanation and the without-explanation condition groups could not have been discovered had existing approaches, in which user trust is measured solely at the conclusion of a task, been employed. Furthermore, the reasons for variations in participants' trust, and variations in the levels of the factors of trust, could not have been understood without capturing their attitudes through the use of the think aloud protocol.

This approach also allows for a greater understanding of the individual factors of trust and the consideration, in the context of intelligent systems, that users may trust the system in terms of one factor (e.g. perceived transparency) but not another factor (e.g. perceived system ability). There is great value in developing and expanding this approach to measure user trust across multiple tasks carried out over a longer period of time, not limited to a single session.

A greater comprehension of not only what engenders trust, but why users trust and how their trust evolves over time - their trust journeys - can lead to the design of more trustworthy intelligent systems.

REFERENCES

1. Bunt, A., Lount, M., and Lauzon, C. Are explanations always important?: A study of deployed, low-cost intelligent interactive systems. In *Proc. IUI 2012*, ACM Press (2012), 169-178.
2. Cramer, H., Evers, V., Ramlal, S., van Someron, M., Rutledge, L, Stash, N., Aroyo, L., and Wielinga, B. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction 18*, 5 (2008), 455-496.
3. Gambetta, D. Can We Trust? In D. Gambetta (Ed.) *Trust: Making and Breaking Cooperative Relations*, Blackwell (1988), 213-237.
4. Glass, A., McGuinness, D.L., and Wolverton, M. Toward establishing user trust in adaptive agents. In *Proc. IUI'08*, ACM Press (2008), 227-236.
5. Hoffmann, A., Söllner, M., Hoffmann, H., and Leimeister, J.M. Towards Trust-Based Software Requirement Patterns. In *Requirements Patterns (RePa), 2012 IEEE Second International Workshop on*, IEEE (2012), 7-11.
6. Höök, K. Steps to take before intelligent systems become real. *Interacting with Computers 12*, 4 (2000), 409-426.
7. Kramer, R.M. Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*, 1 (1999), 569-598.
8. Kulesza, T., Stumpf, S., Burnett, M., Wong, W., Riche, Y., Moore, T., Oberst, I., Shinsel, A., and McIntosh, K. Explanatory Debugging: Supporting End-User Debugging of Machine-Learned Programs. In *Proc. VL/HCC 2010*, IEEE (2010), 41-48.
9. Lewis, J.D., and Weigert, A.J. Trust as a social reality. *Social Forces*, *63*, 4 (1985), 967-985.
10. McKnight, D.H., and Chervany, N.L. Trust and distrust definitions: One bite at a time. In *Proc. Workshop on Deception, Fraud, and Trust in Agent Societies* (2001), 27-54.
11. Muir, B.M. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*, 11 (1994), 1905-1922.
12. Pu, P. and Chen, L. Trust building with explanation interfaces. In *Proc. IUI'06*, ACM Press (2006), 93-100.
13. Rempel, R.E., Holmes, J.G., and Zanna, M.P. Trust in close relationships. *Journal of Personality and Social Psychology*, *49*, 1 (1985), 95-112.
14. Söllner, M., Hoffmann, A., Hoffmann, H., and Leimeister, J.H. How to Use Behavioral Research Insights on Trust for HCI System Design. *Ext. Abstracts CHI 2012*, ACM Press (2012), 1703-1708.
15. Tang, J., Gao, H., Liu, H., and Das Sarma, A. eTrust: Understanding trust evolution in an online world. In *Proc. SIGKDD 2012*, ACM (2012), 253-261.
16. Tintarev, N., and Masthoff, J. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, *22*, 4-5 (2012), 399-439.