Trust in Human-Computer Interactions as Measured by Frustration, Surprise, and Workload

Leanne M. Hirshfield, Stuart H. Hirshfield, Samuel Hincks,
Matthew Russell, Rachel Ward, and Tom Williams

Department of Computer Science, Hamilton College, Clinton, NY 13323, USA
{lhirshfi,shirshfi,shincks,mprussel,rward,tewillia}@hamilton.edu

Abstract. We describe preliminary research that attempts to quantify the level of trust that exists in typical interactions between human users and their computer systems. We describe the cognitive and emotional states that are correlated to trust, and we present preliminary experiments using functional near infrared spectroscopy (fNIRS) and electroencephalography (EEG) to measure these user states. Our long term goal is to run experiments that manipulate users' level of trust in their interactions with the computer and to measure these effects via non-invasive brain measurement.

Keywords: fNIRS, EEG, electroencephalograph, near-infrared spectroscopy, workload, frustration, surprise, trust.

1 Introduction

We describe research that attempts to model and quantify the level of trust[1] that exists in typical interactions between human users and their computer systems. This can be useful for a number of reasons. For example, in order for users to interact with online websites, they must trust the security and validity of that site. If we can measure users' levels of trust during online interactions in usability studies, we can ensure that a given website is designed appropriately to maximize users' trust. Measuring trust would also be useful during usability studies of a number of applications and technologies, pointing designers to areas of the technology or interface that should be re-designed to maximize the user's comfort while working with the system. In addition, measuring trust can help to defer the wealth of money and time spent on training personnel to detect security breaches. If we can measure users' changing levels of trust while working with their computer systems, we can have a better understanding of the training needed to ensure that security personnel detect breaches quickly and accurately.

[1] There is a great deal of research from the management, economics, and recently, from the computer science disciplines, that focuses on building definitions and models that describe the concept of' trust'. While we describe some of this research in section 2, this work does *not* present new models or definitions of trust. We use the term in its most general, non-specific sense, as it is the word choice at this time that best describes the elements of the human-computer interactions that we are exploring.

D.D. Schmorrow and C.M. Fidopiastis (Eds.): FAC 2011, HCII 2011, LNAI 6780, pp. 507–516, 2011.
© Springer-Verlag Berlin Heidelberg 2011

The novelty of our work stems primarily from the multi-modal approach of our analysis, which employs standard psychological testing techniques (both during and post-experiment surveys) to pinpoint emotional and cognitive components of trust as well as brain measuring technologies (EEG and functional near infrared spectroscopy, or fNIRS, recordings) to record physiological reactions. Specifically, our research focuses on the changing relationship from trust to distrust over time from the perspective of a person who is detecting deceitfulness, rather than being deceitful.

In our initial experiment we used a variation of "The Trust Game", a scenario which has been used by many trust researchers. In our version, we modified the level of trustworthiness of the computer agent with whom the subject was playing as the game progressed. Throughout the experiment, we measured, via subjective surveys, the cognitive and emotional changes that occurred while the user's level of trust toward the computer agent changed. The results indicated that the cognitive and emotional states of workload, frustration, and surprise are directly correlated to the users' changing level of trust with a computer agent, and that as users lost trust in the computer agent, they had increasing levels of workload, surprise, and frustration.

We hypothesize that these findings can be generalized and used to measure the amount of trust a user feels during a variety of human-computer interactions. Thus, if we can objectively measure users' levels of workload, surprise, and frustration throughout human-computer interactions, we can understand that user's current level of trust toward the computer, the computer agent, the website, or any other computer mediated communication with which the user is interacting. Our long term goal is to measure changing levels of trust during human-computer interactions. As a first step, we have been running preliminary experiments that isolate the cognitive components of workload, frustration, and surprise.

The rest of this paper is organized as follows: First, we provide a brief overview of research dealing with the user state of trust. Second, we describe the experiment that we conducted to manipulate trust and to understand the cognitive and emotional state changes that were correlated to trust. Third, we provide an overview of the past research on measuring the user states of workload, frustration, and surprise. Fourth, we describe our EEG and fNIRS devices, and we discuss the preliminary experiments that we ran with these devices as well as our experiment results. Fifth, we provide analysis of our results. Lastly, we describe avenues for future work in the measurement of trust during human-computer interactions.

2 Background and Relevant Literature

The topic of trust has sparked a wealth of research in the domains of management and economics. There are many working definitions of 'trust' which have been proposed in the literature[1-5]. The rapid evolution of the internet and the viruses, hackers, and malware that have surfaced in recent years allow for new, broader interpretations of previous notions of interpersonal trust [6]. How much do we trust our computers and the content being presented to us by our computers? What elements of the human-computer interaction affect our feelings of trust toward a computer, computer agent, or web site? Can we 'trust' a computer in the first place? We chose to use the term 'trust' throughout our paper because our research findings are based on the well

known 'Trust Game' which has been used extensively in the trust literature, and because 'trust' is the word that is used most often in qualitative interviews that have been conducted in our lab while people talk about their interactions with computers.

2.1 Trust Experiment

We conducted an experiment that aimed to discover relationships between a computer user's level of trust and that user's changing cognitive and emotional states. To do so we asked participants to sit in front of a standard-size computer monitor and interact with a computer console on the computer screen. They then played a version of the "Trust Game," developed by Berg [4], which has been used in many experiments dealing with trust, risk-taking, and money management [3]. In our version of the "Trust Game" both the computer and the user began with a fictional $10. The user and computer would take turns sending some amount of money, ranging from $0-$10, back and forth to one another. Each time some amount of money was sent between the user and computer, the amount sent was tripled while en route. In an ideal, high-trust scenario, the computer and the user would always send a the maximum amount of money back and forth to one another—maximizing the gain possible for each of the two Trust Game 'players'. This process was repeated 23 times.

For the first eight transactions, the computer acted "trustworthy" in the sense that it typically returned a high amount of money back to the participant. For the next nine transactions, the computer acted with a mix of trustworthy and untrustworthy behavior, sometimes returning a high amount, and other times returning a low amount of money back to the participant. For the last six transactions the computer acted in a wholly untrustworthy manner, regularly returning a very low amount of money back to the participant.

We used data from Self Assessment Manikins which were administered after every six transactions to gauge the user's cognitive and emotional state. Additionally, we collected the amount given and percentage returned for each transaction, as well as information regarding the subjects' self-rated locus of control, trust, and computer familiarity.

From this information we gleaned that the median amount given increased during the computer's trustworthy state, varied highly during the computer's erratic state, and decreased during the computer's untrustworthy state. Furthermore, we analyzed the data from the Self Assessment Manikins, turning users self reported measures of valence, arousal, and dominance into a set of discrete user states. We found a direct correlation between the trustworthiness of the computer agent and the user's reported measures of workload, frustration, and surprise. Our results showed that overall workload increased throughout the experiment, as did frustration and surprise.

We were also interested in whether relationships existed between frustration, surprise, and workload. We found that significant correlations between frustration and workload in all four surveys existed. We also found significant correlations between workload and surprise in surveys 1, 3, and 4, as shown in Table 1 below.

Our results suggested that as computer users lost trust with the simulated agent they were interacting with in the Trust Game, the user states of workload, surprise, and frustration were directly correlated with the users' changing level of trust.

Table 1. Significant Correlations between workload/frustration, and workload/surprise

	Survey 1	Survey 2	Survey 3	Survey 4
Frustration & Workload	$r(27)= .565$ $p<.01$	$r(25)=.733$ $p<.01$	$r(25)=.67$ $p<.01$	$r(26)=.739$ $p<.01$
Workload and Surprise	$r(27)=.559$ $p<.01$	$r(25)=.391$ $p<.06$	$r(25)=.638$ $p<.01$	$r(26)=.478$ $p<.02$

2.2 Linking Trust in Human-Computer Interactions to Surprise, Workload, and Frustration

While these results are restricted to the version of the Trust Game that users played, we hypothesize that the same user states will influence the level of trust in more realistic interactions between users and their computer systems. As an illustrative example, consider John Doe's interaction with his computer over the course of a year:

When John's computer was functioning properly, he had high trust in his interactions with and through his computer. During these high trust times, John had low levels of frustration, workload, and surprise—all interactions with his computer seemed to proceed as expected. However, one day John visited a new website and suddenly he noticed hundreds of pop ups infiltrating his screen (i.e., surprise and frustration). He later found out that his computer had a virus, which likely came from the site with all of the pop ups (frustration). Later on that year, John was Instant Messaging with his friend Alice. Alice was a classmate of John's and a user with her name as a username contacted John via IM. After a few minutes of messaging with who he presumed to be Alice, John began to become wary of the interactions. The IMer was not writing in a way that was consistent with Alice. John began to interact very cautiously with the IMer, hoping to determine if the person was indeed an imposter (workload, frustration, surprise). Also, over the course of time, John's computer became very slow because he downloaded too many programs and add-ons. He was frustrated while using it because it took a long time for him to get things done and he found it difficult to keep focused on the task at hand while waiting long intervals for his computer to catch up to his train of thought (frustration and workload).

All of these occurrences caused John's level of trust during his computer interactions to be lowered. We hypothesize that we can use measures of users' workload, frustration, and surprise to indicate that users' level of trust. In the next sections we describe cutting edge research that attempts to measure these user states objectively.

2.3 Measurement of Surprise, Workload, and Frustration

Acquiring quantitative data about computer users is a continual challenge for researchers in HCI. Although we can accurately measure task completion time and accuracy, measuring factors such as mental workload, frustration, and distraction are often done by qualitatively observing users or by administering subjective surveys

to users. These surveys are often taken after the completion of a task, potentially missing valuable insight into the user's changing experiences throughout the task. They also fail to capture internal details of the operator's mental state. To address these evaluation issues, much current research focuses on developing objective techniques to measure, in real time, user states such as workload, frustration, and surprise [7-9]. Although this ongoing research has advanced user experience measurements in the HCI field, finding accurate and non-invasive tools to measure computer users' states in real working conditions remains a challenge. The user states that are addressed by this research are the states of workload, frustration, and surprise.

Surprise. Surprise has previously been measured in HCI studies using facial analysis software [10] as well as using Skin Conductivity, blood volume and heart rate [11]. Detecting surprise with electroencephalography (EEG) is a topic of much research in Psychophysiology. Surprise can be indicated by the presence of an Error-Related Potential (ErrP), in which EEG data contains error-related negativity (a sharp negative deflection around 80ms)[12], often followed by error-related positivity (a slow positive wave)[13]. ErrPs can be found both when a user makes an error and when a user notices an error made by the machine [14]. This makes the measurement of ErrPs useful in HCI-based interface analysis.

Frustration. Frustration is an important metric in HCI. Lazar, et al. studied frustration with computer interfaces by having employees that used computers in their workday keep journals that tracked their ongoing frustration as they used their routine computer programs [15]. The results showed that word processing and email were reported as the most frustrating activities, and that participants wasted an average of forty percent of their time trying to solve unnecessarily frustrating problems.

Biological methods of measuring frustration allow researchers to collect more reliable data. Scheier, et al., induced frustration in users with a mouse that sporadically froze and inhibited the user from winning a game [16]. A Hidden Markov Model analyzed skin conductivity, blood volume pressure and the state of the mouse, eventually learning the manifestations of frustration. The results indicate that a user's affective state can be automatically discriminated from events in their physiology [16]. Most recently, BCI devices enable automated detection of user frustration through discovering patterns in brain activity. Reuderink et al. developed an affective version of Pacman which places the user in a state of frustration [8]. They found significant differences in EEG activity during periods of frustration and the normal state.

Workload. The ability to acquire objective, real-time measures of a computer user's mental workload while (s)he works with a computer would be valuable to the field of HCI. Adaptive interfaces could adapt in real-time to a given user based on his or her current level of workload, keeping that user in *the flow* [17]. Also, measures of users' mental workload could be acquired during usability studies to help interface designers to pinpoint areas of the interface that may be un-intuitive for users [18, 19].Researchers have successfully used EEG or fNIRS to measure elements of mental workload such as working memory[20-23], response inhibition [21, 24], visual search [21, 25], as well as a myriad of other executive processes [26, 27].

3 Measurement Oriented Experiments

We conducted three preliminary experiments where we attempted to manipulate and measure the user states of surprise, frustration, and workload. In the future, we aim to acquire real time measures of these user states in order to predict one's level of trust during his or her computer interactions.

3.1 Surprise Experiment

An important component of trust is the moment of surprise; that is the moment when a person notices that something 'unexpected' has occurred in the computer system. This could be the moment users notice that a virus is on their computer, or the moment they realize that the person they are IMing with may be an imposter. To measure this, we exploited the oddball paradigm in order to elicit surprise. Three participants completed an experiment that was created using Eprime in which they pressed two different buttons depending on the position of an oval on the screen. The oval was in one of two positions, located either on the far left or the far right side of the screen. When the oval was on the left side of the screen the subjects were instructed to press the 'z' button, and when the oval was on the right side of the screen they were instructed to press the 'm' button.

Immediately following the subject response a feedback screen indicated whether or not the subject had pressed the correct key. Subjects completed 150 tasks where they simply hit the 'z' or 'm' keys to indicate the position of the oval on the screen. During the first 20 tasks, the feedback for the subjects was as expected. During the last 130 tasks, we randomly selected 15% of the tasks to provide incorrect, *or surprising feedback* to the user. In other words, 15% of the time, when subjects pressed the 'z' key, the feedback indicated that the 'm' key had been pressed, and vice versa.

The EEG used in the study was Advanced Brain Monitoring's b-alert wireless 10 channel EEG. Data was sampled at 256Hz (www.b-alert.com). The non-invasive EEG is an ideal brain monitoring device for use in human-computer interaction studies, where it may be important to keep participants comfortable while completing tasks in realistic working conditions.

The Eprime software sent markers to the EEG immediately before the subject saw the feedback screen. In this way, we planned to search for the presence of an ErrP that was caused when the surprising feedback occurred during 15% of the tasks.

Data Analysis and Results of Surprise Experiment. We used a similar procedure as Ferrez et al. [14] to preprocess our EEG data for classification. We took the data from the moment the feedback occurred through to 650ms after the feedback was shown for channels Cz and Fz. Like Ferrez et al., we chose these channels because ErrPs are usually found in a fronto-central distribution along the midline [14] Each temporal section of data was associated with one of two class labels: control or surprise, indicating whether or not the feedback the subject saw at that moment was the expected feedback or the surprising feedback.. We applied a 1-10 Hz bandpass filter as ErrPs have a relatively slow cortical potential. We downsampled our data from 256Hz to 128Hz and input our resulting timeseries data into a weighted K-nearest neighbor classifier (k = 3) with a Dynamic Time Warping distance measure. We ran

our classification separately for each subject. Results are in Table 2. We were able to distinguish between the control and surprising feedback conditions with an average of 71% accuracy for our three subjects.

Table 2. Classifier accuracy distinguishing between the control and surprising feedback

	sub1	sub2	sub3	average
Classifier Accuracy	70%	74%	68%	71%

3.2 Frustration Experiment

In this section we report on an experiment that was completed in 2009 [28]. During the experiment six subjects completed a series of nback tasks [20, 21], which have been used in many experiments to manipulate working memory. In the 1-back task, depicted in Figure 1, subjects must indicate whether the current letter on their computer screen is a match ('m'), or not a match ('n') to the letter that was shown 1 screen previously.

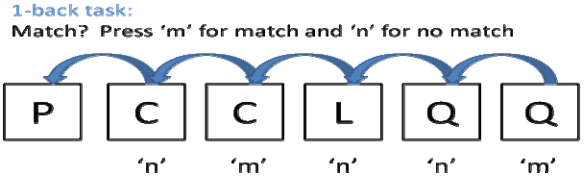

Fig. 1. Depiction of the 1 back task

Each task lasted 30 seconds with a rest time of 20 seconds between tasks. Half of the 1back tasks were completed by subjects as expected. However, during the other half of the 1back tasks, internet pop ups such as the one shown in Figure 2, were introduced into the computer systems. Subjects were told to finish the nback tasks as quickly as possible and with the highest accuracy possible. Six subjects (3 female, 3 male) completed the experiment. Subjects were all Tufts undergraduate students. A randomized block design with eight trials was used in this experiment.

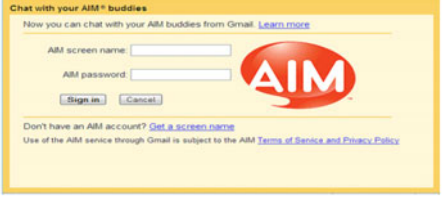

Fig. 2. An example of a pop up in the frustration experiment

In this experiment we used an OxyplexTS (ISS Inc. Champagne, IL) frequency-domain tissue spectrometer with two optical probes. Each probe has a detector and four light sources. Each light source emits near infrared light at two separate wavelengths (690nm and 830nm) which are pulsed intermittently in time. This results in 2 probes x 4 light sources x 2 wavelengths = 16 light readings at each timepoint (sampled at 6.25Hz).

Data Analysis and Results of Frustration Experiment. All subjects were interviewed following the experiment. All subjects indicated that the pop ups were a source of frustration throughout the experiment. We computed all machine learning analyses separately for each subject. For each subject, we recorded 16 channel readings throughout the experiment where we refer to the readings of one source detector pair at one wavelength, as one *channel*. We normalized the intensity data in each channel by their own baseline values. We then applied a moving average band pass filter to each channel (with values of .1 and .01 Hz) and we use the modified Beer-Lambert Law[12] to convert our light intensity data to measures of the relative changes in oxygenated (HbO) and deoxygenated hemoglobin (Hb) concentrations in the brain. This resulted in eight readings of HbO and eight readings of Hb data at each timepoint in the experiment. We then averaged together the channels from the left side of the head and the channels on the right side of the head, giving us 4 time series for each subject; 1) HbO on the left side of the head, 2) HbO on the right side of the head, 3) Hb on the left side of the head, and 4) Hb on the right side of the head. We then input these time series into a weighted KNN classifier (k = 3) with a distance measure computed via Symbolic Aggregate Approximation (SAX). For more information on SAX, see [29]. As shown in Table 3, we were able to distinguish between the control 1back tasks and the frustrating 1back tasks with an average of 73% accuracy across the six subjects.

Table 3. Classifier accuracy at distinguishing between the control (1back) and frustrating (1back with pop-ups) conditions

	sub1	sub2	sub3	sub4	sub5	sub6	average
Classifier Accuracy	69%	81%	63%	75%	75%	75%	73%

3.3 Workload Experiments

We have conducted several experiments, using the fNIRs device described above, to measure various aspects of mental workload. Using this device we have:

1. Used machine learning techniques to classify, on a single trial basis, the load placed on users visual search, working memory, and response inhibition resources [21].
2. Used machine learning techniques to classify various levels of working memory load in a simple counting and addition task [30].
3. Used machine learning techniques to distinguish between spatial and verbal working memory [19].

4 Conclusion and Future Work

We described preliminary research that attempts to quantify the level of trust that exists in typical interactions between human users and their computer systems. We described the cognitive and emotional states that we found to be correlated to trust, and we presented preliminary experiments using functional near infrared spectroscopy and electroencephalography to measure these user states. The experiments presented in this paper represent the beginning of our research on the measurement of trust during human-computer interactions. Ongoing work in our lab continues to manipulate, and measure, the user states of frustration, surprise, and workload. The longer term goal is to run experiments that manipulate users' level of trust in their interactions with the computer and to measure these effects via non-invasive brain measurement.

References

1. Mayer, R., et al.: An Integrative Model of Organizational Trust. The Academy of Management Review 20(3), 709–734 (1995)
2. Serva, M.A., Fuller, M.A., Mayer, R.: Trust in systems development: a model of management and developer interaction research in progress. In: Proceedings of the 2000 ACM SIGCPR Conference on Computer Personnel Research, ACM, Chicago (2000)
3. Lewicki, R., et al.: Trust and Distrust: New Relationships and Realities. The Academy of Management Review 23(3), 438–458 (1998)
4. Berg, J., Dickhaut, J., McCabe, K.: Trust, Reciprocity, and Social History. Games and Economic Behavior 10, 122–142 (1995)
5. McAllister, D.J.: Affect- and cognition-based trust as foundations for interpersonal cooperation in organizations. Academy of Management Journal 38(1) (1995)
6. Schneider, F.: Trust in Cyberspace. National Academy Press, Washington (1998)
7. Mandryk, R., Atkins, M., Inkpen, K.: A continuous and objective evaluation of emotional experience with interactive play environments. In: Proceedings of the SIGCHI conference, ACM Press, Montreal Canada (2006)
8. Reuderink, B., Nijholt, A., Poel, M.: Affective Pacman: A Frustrating Game for Brain-Computer Interface Experiments. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering (2009)
9. Savran, A., et al.: Emotion Detection in the Loop from Brain Signals and Facial Images. In: eNTERFACE 2006, Dubrovnik, Croatia (2006)
10. Ward, R.: An analysis of facial movement tracking in ordinary human-computer interaction. Physiological Computing 16(5), 879–889 (2004)
11. Ward, R., Marsden, P.: Physiological responses to different web page designs. International Journal of Human Computer Studies 59, 199–212 (2003)
12. Chavarriaga, R., Ferrez, P., Millán, J.: To Err is Human: Learning from Error Potentials in Brain-Computer Interfaces. Advances in Cognitive Neurodynamics, 777–782 (2008)
13. Nieuwenhuis, S., et al.: Psychophysiology, Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task: p. 752–760
14. Ferrez, P., Millán, J.: You Are Wrong!—Automatic Detection of Interaction Errors from Brain Waves. In: Proceedings of the 19th International Joint Conference on Artificial Intelligence (2005)
15. Lazar, J., Jones, A.S.: Workplace user frustration with computers: an exploratory investigation of the causes and severity. Behaviour & Information Technology, 239–251 (2006)

16. Scheirer, J., et al.: Frustrating the user on purpose: a step toward building an affective computer. Interacting with Computers, 93–118 (2002)
17. Csikszentmihalyi, M.: Flow: The Psychology of Optimal Experience, Harper Collins, 320 (1991)
18. Lee, J.C., Tan, D.S.: Using a low-cost electroencephalograph for task classification in HCI research. In: Proceedings of the 19th annual ACM symposium on User interface software and technology, ACM Press, Montreux (2006)
19. Hirshfield, L.M., et al.: Brain Measurement for Usability Testing and Adaptive Interfaces: An Example of Uncovering Syntactic Workload in the Brain Using Functional Near Infrared Spectroscopy. In: Conference on Human Factors in Computing Systems: Proceeding of the twenty-seventh annual SIGCHI conference on Human factors in computing systems (2009)
20. Grimes, D., et al.: Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph. In: CHI 2008 Conference on Human Factors in Computing Systems, Florence, Italy (2008)
21. Hirshfield, L., et al.: This is your brain on interfaces: enhancing usability testing with functional near infrared spectroscopy. In: SIGCHI, ACM, New York (in press, 2011)
22. Sassaroli, A., et al.: Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy (2009); accepted in the Journal of Innovative Optical Health Sciences
23. Gevins, A., et al.: High-Resolution EEG Mapping of Cortical Activation Related to Working Memory: Effects of Task Difficulty, Type of Processing, and Practice. Cerebral Cortex (1997)
24. Schroeter, M.L., et al.: Near-Infrared Spectroscopy Can Detect Brain Activity During a Color-Word Matching Stroop Task in an Event-Related Design. Human Brain Mapping 17(1), 61–71 (2002)
25. Anderson, E.J., et al.: Involvement of prefrontal cortex in visual search. Experimental Brain Research 180(2), 289–302 (2007)
26. Tanida, M., et al.: Relation between asymmetry of prefrontal cortex activities and the autonomic nervous system during a mental arithmetic task: near infrared spectroscopy study. Neuroscience Letters 369(1), 69–74 (2004)
27. Joanette, Y., et al.: Neuroimaging investigation of executive functions: evidence from fNIRS. PSICO 39(3) (2008)
28. Hirshfield, L.M.: Enhancing Usabiltiy Testing with Functional Near Infrared Spectroscopy. In: Computer Science, Tufts University, Medford (2009)
29. Lin, J., et al.: A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In: Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, San Diego, CA (2003)
30. Hirshfield, L.M., et al.: Human-Computer Interaction and Brain Measurement Using Functional Near-Infrared Spectroscopy. In: Symposium on User Interface Software and Technology: Poster Paper, ACM Press, New York (2007)