Investigating Trust in Automation

Hailee Aklyan

Masters Lab Research

University of Idaho

Investigating Trust in Automation

Abstract

Automated systems have become more prominent in human-computer interaction with the issue of trust at its height. This study aimed to examine the repercussions of user trust when automation errors occur. Participants were asked to play a computer multitasking game that adds new tasks at regular intervals of 15 seconds, 40 seconds, and 80 seconds. There was a total of four trials; trial 2 and trial 4 had researcher-induced automation. Automation in trial 2 would perform accurately while automation in trial 4 would perform with error. A significant difference between the means was indicated by the data with an effect size of 98%. The limitations of this study will be useful for addressing future studies in this field.

Keywords: automation, level of trust, multitasking game, automation error, task

Investigating Trust in Automation

Automation is defined as the performance by machines of functions that were previously carried out, whether fully or partially, by humans. Automation has been used to perform tasks that humans are incapable of doing but also has been used to assist users. There are many benefits to using automated systems like precision of performance, increased productivity, and enhanced safety. Although there are many benefits, there are also drawbacks. Automation technology is not 100% sound and has potential flaws. Accidents related to automation have occurred and have the potential to affect a user's performance negatively.

Trust plays a large role in the success of automation technology. It comes from the cognitive state of the operator/user and is subjective in nature. Trust can be quantified for research purposes with questionnaires. There are three different broad levels of trust: calibrated trust, over-trust, and under-trust. Users can fall into any one of these categories and is dependent on their reliability in the automation. Many research opportunities for automated system design present themselves, however, unless a mitigation strategy for increasing user trust is supported, the success of automated systems will falter.

Study Aim and Hypothesis

This study aimed to examine the repercussions of user trust when automation errors occur. I was interested in changes in a user's subjective level of trust following error in an automated task. Participants were asked to play a visual-spatial multitasking game as the primary task. The game began with a single task and additional tasks were added at regular intervals. The game kept track of time which I used as a measure of performance. If users trust the efficacy of an automated system, then I will observe a change in trust after an error occurs.

Methods

Investigating Trust in Automation

Sample

Participants were recruited from a population of students aged 18-25 enrolled in undergraduate psychology courses at the University of Idaho. I studied a total of 20 participants as a participant pool of this size should provide sufficient data to test my hypotheses. I was not concerned with any participant characteristics other than severe visual or auditory impairments, which were the only exclusion criteria.

Measures

Prior to phase 1, basic participant demographics were collected. Testing in Phase 2 was completed by measuring final scores, the number of games reached, and a subjective measure of trust. I measured trust in automation based on a quantitative survey developed by Jiun-Yin et al. (see Appendix). In this survey, trust in automation was based on calculating the sum of participants' ratings for each question. The minimum score is 12 and the maximum score is 84; the higher the score, the more a participant trusts the automation and vice versa. This survey was taken at the conclusion of trial 2 and trial 4.

Procedure

The multitasking game will remain the constant primary task throughout the duration of the study. The study followed an ABAB design. I used this design to evaluate the strength of the treatment variables and the relationship they had with the dependent variable. Additionally, the ABAB design allowed comparison between trials in different phases of testing.

Phase 1:

First, the informed consent process was initiated, and basic demographics were collected. The participant was asked if they had any questions or needed clarification on anything in the

informed consent. The participant was asked to play through the first task and was asked to stop when the next task was added for further instruction.

Phase 2:

For Trials 1 and 3, participants were tasked with manually playing the multitasking game. In these trials, the participants were told that they are not provided with any assistance in playing the multitasking game. I was looking at measuring performance without the presence of assistive automation. I recorded the participant's final score and the number of games reached.

Trial 2, was the first time participants were exposed to “automation.” They were told that the automation would assist with the first game or task. My research partner used a secondary keyboard to mimic the “automation” of the first game. In this trial, the participant did not experience any errors in automation. As a result of reliability, some level of trust in automation occurred. At the conclusion of this trial, I surveyed the participant's current level of trust in the game’s automation using a trust-in automation questionnaire. I also recorded the participant's final score and the number of games reached.

In Trial 4, participants were notified of the presence of automation assisting in the same way as in Trial 2. However, in this trial, the “automation” experienced errors. Therefore, it was up to the participant to detect these errors and either correct them or fail. Afterward, I recorded the final score, the number of the participant’s subjective measure of trust in the game’s “automation.”

Roadmap



Results

At the conclusion of the data collection, a paired-sample t-test was conducted to compare the overall trust questionnaire score for trial 2 and trial 4. Results showed a significant difference between the trial 2 questionnaire score (M=48.80, SD=4.99) and the trial 4 questionnaire score (M=40.25, SD=8.84); [t(19)=4.36, p=<0.001]. The data indicates there was a significant difference between the means with a 95% confidence interval and an effect size of 98%.

Investigating Trust in Automation

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Trial 2 Questionairre Score	48.80	20	4.991	1.116
	Trial 4 Questionairre Score	40.25	20	8.837	1.976

Paired Samples Correlations

		N	Correlation	Significance One-Sided p	Significance Two-Sided p
Pair 1	Trial 2 Questionairre Score & Trial 4 Questionairre Score	20	.297	.102	.203

Paired Samples Test

		Paired Differences						
					95% Confidence ...		Significance	
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper	One-Sided p	Two-Sided p
Pair 1	Trial 2 Questionairre Score - Trial 4 Questionairre Score	8.550	8.763	1.959	4.449	12.651	4.364	19	<.001	<.001

Paired Samples Effect Sizes

			Standardizer[a]	Point Estimate	95% ... Lower	95% ... Upper
Pair 1	Trial 2 Questionairre Score - Trial 4 Questionairre Score	Cohen's d	8.763	.976	.431	1.503
		Hedges' correction	8.941	.956	.423	1.473

Discussion

The goal of this study was to examine the repercussions of user trust when automation errors occur. If the hypothesis was correct, inducing automation error would decrease a user's level of trust. The data indicated a significant difference with a 95% confidence interval; therefore, I must reject the null hypothesis that inducing automation errors will not affect a user's level of trust in automation.

Limitations

Although the data of this study identified significant results, there are still some limitations that must be addressed. First, the terminology and verbiage used on the questionnaire were not comprehensible to some participants. They needed further clarification and simpler definitions for words like “deceptive,” “wary,” and “underhanded.” For future studies, a simpler vocabulary could be used to better clarify to participants the meaning and decrease questionnaire misunderstanding. There is a chance that the misunderstanding led participants to answer a question in a different manner. A second limitation was the lack of protocol depth; I did not expect to be asked questions about how the automation worked and if the user needed to still control it. This decreased my internal validity because I did not write a script with answers to these questions, therefore, users did not get the same instruction or answer. Future studies must address this in the protocol and script prior to beginning their study to decrease the threat to internal validity. Finally, the sample size (N=20) was small and potentially affected the power of the study. Increasing the sample size in the future would increase the power of the study and may lead to more potentially significant results. Overall, further research in the field of automation could acquire essential information to mitigate the current lack of trust there is in automation. As newer products are available with automation, users’ level of trust will have a large effect on the success of the product. Automation has been found to decrease mental workload when used as an aid for users and has also been seen to benefit users in task allocation.