Automation reliability and trust: A Bayesian inference approach

Chenlan Wang[1], Chongjie Zhang[2], X. Jessie Yang[1]
1. Industrial & Operations Engineering Department
University of Michigan, Ann Arbor, MI
2. Institute for Interdisciplinary Information Sciences
Tsinghua University, Beijing, China

Research shows that over repeated interactions with automation, human operators are able to learn how reliable the automation is and update their trust in automation. The goal of the present study is to investigate if this learning and inference process approximately follow the principle of Bayesian probabilistic inference. First, we applied Bayesian inference to estimate human operators' perceived system reliability and found high correlations between the Bayesian estimates and the perceived reliability for the majority of the participants. We then correlated the Bayesian estimates with human operators' reported trust and found moderate correlations for a large portion of the participants. Our results suggest that human operators' learning and inference process for automation reliability can be approximated by Bayesian inference.

INTRODUCTION

Today automation is assisting human operators in many areas including military tasks, healthcare systems, operations in factories, and public transportation systems. Although automated technologies are becoming increasingly more intelligent and reliable, there are unexpected situations automation cannot handle. The human operators are, therefore, required to resume control of the task. In order for the human-automation team to work optimally, it is important for the human operators to understand the reliability of the automation and to calibrate their trust in automation appropriately.

Trust in automation is defined as the attitude that individuals' goals can be assisted to achieve by an agent in situations with uncertainty and vulnerability (Lee & See, 2004). Existing research on trust mainly treat trust as a steady-state variable and measure trust in automation via questionnaires administered at the end of an experiment. There is limited amount of research on the dynamic nature of trust (Lee & Moray, 1992; Desai et al., 2012; Khasawneh et al., 2003; Manzey et al., 2012; Yang et al. 2017).

Studies examining trust dynamics suggest that with repeated interactions with automation, human operators are able to learn its system reliability overtime and continuously update their trust. In the present study, our goal is to investigate if this learning and inference process approximately follow the principles of Bayesian probabilistic inference. Using an existing dataset, we first applied Bayesian probabilistic inference to estimate human operators' perceived automation reliability. We then compared the Bayesian estimates against the values reported by the human operators and found high correlations between the two. At last, we correlated the Bayesian estimates with the human operators' reported trust and found moderate correlations. Our results suggest that human operators' learning and inference process of automation reliability can be approximated by Bayesian inference.

RELATED WORK

Trust in automation has received a substantial amount of attention in the past two decades and factors influencing trust in automation have been well documented (see Hoff & Bashir, 2015; Schaefer et al., 2016 for reviews). However, the majority of existing research treated trust as a steady-state variable. There were few studies examining how trust changes as operators gain more experience with the automation.

Trust is dynamic, and it can increase or decay overtime (Yang et al. 2017). Lee & Moray (1992) studied the change of trust in a pasteurization controlling task. They proposed a time series model ARMAV (An autoregressive moving average vector form) to capture the dynamics of trust. According to the ARMAV model, trust at the present moment is determined by trust at the previous moment, automation performance, and the occurrence of automation failures. Desai et al. (2012) studied the influence of system reliability drops at the beginning, middle, or end of the designed path in a robot controlling tasks. Their results showed that the participants' trust was influenced by the change of reliability and was influenced more strongly when the automation failures occurred near the end than at the start or in the middle.

Along the same line, Khasawneh et al. (2003) developed a multiple regression model of trust for a hybrid inspection system. Participants were required to observe the computer performance of a visual inspection task. They were informed of the true state of the task each trial and indicated their trust. Results of the study showed that human trust depended on the automation performance and the true state of the printed circuit boards for the inspection task. Human trust reached the maximum when the computer inspection result matched the actual state of the board and decreased when not matched. Manzey et al. (2012) investigated the dynamics of trust in a supervisory control task. Their study showed that trust was changing with the participants' negative and positive experience. Negative experience had more influence on trust than positive experience and a single automation failure led to immediately decrease of trust. Using a tracking and detection task, Yang et al. (2017) examined how trust in automation

evolved as human operators gained more experience and modeled the evolution process using a first-order linear time invariant dynamic system. Results of their study showed that human operators' trust in automation would stabilize over repeated interaction with an automated technology. The stabilized trust was higher when the system had a higher system reliability.

The above-mentioned studies suggest that over repeated interactions with automation, human operators are able to learn its system reliability overtime and continuously update their trust. In the present study, we aimed to investigate if this learning and inference process approximately follow the principles of Bayesian probabilistic inference.

Bayesian principles dictate how rational agents should update their beliefs in light of new data, based on a set of prior knowledge on the problems at hand (Griffiths, Kemp & Tenenbaum, 2008). Bayesian models have been applied in animal learning (Courville, Daw, & Touretzky, 2006), human inductive learning and generalization (Tenenbaum, Griffiths, & Kemp, 2006, Griffiths et al., 2008), and social cognition (Baker, Tenenbaum, & Saxe, 2007).

DATASET

This paper uses the dataset collected by Yang et al. (2017) (please refer to it for details). Participants in the experiment performed a simulated surveillance task consisting of a tracking task and a detection task (Figure 1). For the tracking task, participants controlled a joystick and moved the green circle to the center of the display as close as possible. Meanwhile, participants detected whether there was a potential threat in four images. Participants were able to access only one task at any time and had to switch between the tracking task and the detection task.

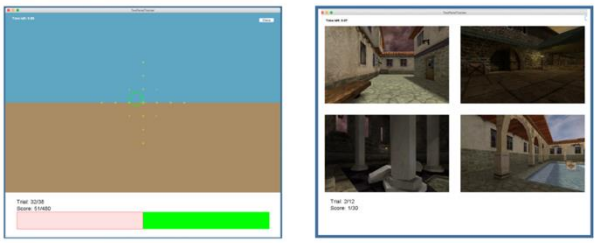

Figure 1. Dual task environment in the simulation testbed (Yang, 2017).

Table 1. Possible states according to SDT

Threat Detector	State of the world	
	Threat	No threat
Danger/Warning	Hit	False Alarm
Clear/Possibly Clear	Miss	Correct Rejection

There was an imperfect threat detector to assist human operators in detecting the threat. The system reliability of the threat detector was set as 70%, 80%, and 90% according to the signal detection theory (SDT). There were four states considering the detection results of the automatic detector and the real states in the world shown in Table 1. The numbers of Hits, Correct rejections, Misses, and False alarms were shown in Table 2.

Table 2. The numbers of Hits, False Alarms (FA), Correct Rejections (CR) and Misses for the three reliability levels

System Reliability	Numbers of each states			
	Hit	Correct Rejection	Miss	False Alarm
~70%	9	59	21	11
~80%	21	59	9	11
~90%	29	59	1	11

Each participant had 100 trials with each trial lasting 10 seconds. After each trial, participants reported their perceived automation reliability, trust in automation, and confidence.

BAYESIAN PROBABILISTIC INFERENCE

This section introduces the mathematics of Bayesian probabilistic inference. The goal is to infer the system reliability (r), a parameter between 0 and 1, by observing a sequence of automation outcomes overtime. For example, imaging that you observed a sequence of automation outcomes TTTTTTTTTT, with T indicating automation success. What is your estimate of the system reliability? How would your belief change if you had observed TTFTFTTFTTF, with T indicating automaton success and F automaton failure?

To formalize this problem in Bayesian probabilistic inference, we need to identify the prior probability of system reliability, *P(r)*, and the likelihood probability of the observed data under the system reliability, *P(d|r)*. At each trial, the automation outcome can either be a success (hit and CR) or a failure (FA and miss), which can be modeled as a Bernoulli distribution. Therefore, the likelihood probability of a particular sequence of automation outcomes containing N_T successes and N_F failures with system reliability r follows a binomial distribution and is given by:

$$P(d|r) = r^{N_T}(1-r)^{N_F}$$

Then we apply Bayes' rule to estimate the system reliability given a sequence of automation outcomes *d*.

$$P(r|d) = \frac{P(d|r)P(r)}{P(d)} = \frac{P(d|r)P(r)}{\int_0^1 P(d|r)P(r)dr}$$

Another factor we need to consider is the strength of human operators' prior knowledge. For example, suppose two operators with the same prior belief that the automaton is 70% reliable. The first operator may not have much prior experience with automated image detection, thus his prior belief is weak. The second operator may have knowledge in imaging processing and his prior belief of the system reliability could be very strong.

We use a Beta distribution to quantify a human operator' prior belief, which is given by

$$r \sim Beta(M_T + 1, M_F + 1)$$

where M_T and M_F are parameters to model mathematically how strong the prior belief is. Using the same example mentioned above, the first participant could be modeled using $M_T = 7$ and $M_F = 3$, and the second participant's strong prior belief could be represented using $M_T = 700$ and $M_F = 300$.

As the class of beta prior distributions is conjugate to the class of binomial likelihood functions, a posterior belief is also a Beta distribution, which can be derived as the following (Griffith et al, 2008):

$$r|d \sim Beta(N_T + M_T + 1, N_F + M_F + 1)$$

System reliability can then be estimated by the posterior mean of the above Beta distribution, denoted as r^*.

$$r^* = \frac{N_T + M_T + 1}{N_T + N_F + M_T + M_F + 2}$$

To support online learning, we use the posterior of the current time step as the prior for the next time step. The online estimation of system reliability was summarized as the algorithm shown in Algorithm 1. The initial prior of the system reliability is a normal distribution, denoted as r_0^*. The total number for prior and the number of trials for each update were denoted as m and n respectively. It is obvious to show that $N_T + N_F = n$ and $M_T + M_F = m$. The total number of automation success in prior at time k (M_{T_k}) was approximated by the prior probability (r_{k-1}^*) and the total number of prior (m). The present system reliability as time k is then calculated using the posterior mean of the Beta distribution mentioned above. The automation performance at the past was added to the total number of prior at the present time.

Algorithm 1. Bayesian Inference Algorithm

$r_0^* = rand(1)$;
For k = 1: n: total number of trials
  $M_{T_k} = r_{k-1}^* m$;
  $r_k^* = \frac{N_{T_k} + M_{T_k} + 1}{n + m + 2}$
  $m = m + n$;
End

RESULTS AND DISCUSSIONS

System Reliability Inference

The dataset used in our study involves automation reliability of 70%, 80%, and 90%, respectively. At each reliability condition, there were twenty-six participants. For every participant worked with the automated threat detector of the same system reliability, the number of automation successes and automation failures were fixed but the sequences were random. We compared the Bayesian's estimate of system reliability against the reported system reliability and calculated the Root Mean Square Error (RMSE) and the Correlation Coefficient (ρ). We set the prior total number m as a variable {5, 10, 20, 50, and 100} to indicate how strong the prior could be (5 indicating a weak prior belief and 100 indicating a strong prior belief). In the Bayesian inference, we also changed the interval n {1, 2, 3, 4, 5, 7, 10, 15, 20} indicating how often a participant could update his belief of automation reliability. For instance, a participant may update his belief after every trial and another may update his belief after every 10 trials.

For each participant, we calculated the optimal (n, m). The corresponding Root Mean Square Error (RMSE) and Correlation Coefficient (ρ) are shown in Table 3.

To a large extent, we observed high correlation and low RMSE between the Bayesian estimate and participants' reported system reliability, suggesting that the learning and inference process approximately follow the principles of Bayesian probabilistic inference (Table 3).

Table 3. Average Correlation Coefficients between Bayesian estimate and participants' reported system reliability

System Reliability	70%	80%	90%	all
Average Correlation $\bar{\rho}$	0.71	0.81	0.82	0.78

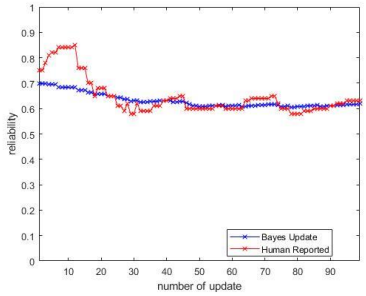

Figure 2. Estimated and Perceived System reliability of Participant 40

To examine the data further, we found that participants could be categorized into three types. In the first type, Bayesian's estimate using the optimal *(n, m)* fits the reported data smoothly. One example is participant #40 (Figure 2), where the Bayesian estimates predicted the participants' reported system reliability accurately. The second type of participants considered the automation as unreliable no matter how the automation performance changed over time. An illustration is shown in Figure 3, where the reported system reliability never exceeded 50%. The third type of participants were sensitive and easily influenced by the automation

performance. An example of the third type is shown in Figure 4. Bayesian method does not work well with the second and the third type of participants even with optimal (n, m).

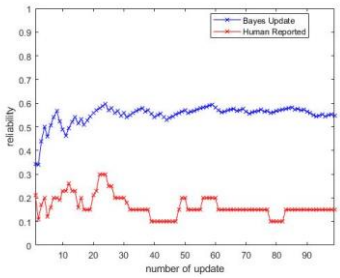

Figure 3. Estimated and Perceived System reliability of Participant 3

Correlation between Estimated and Trust

We calculated the correlation between the Bayesian estimate of system reliability and trust for each participant (Table 4). The results indicate a good correlation between Bayesian estimates of system reliability and human operators' reported trust, with the average correlation coefficient of 0.42.

Moreover, for about half of the participants there was strong correlation ($\rho > 0.7$) between the Bayesian estimates of system reliability and trust, suggest that the pattern of trust change could also be predicted using Bayesian inference.

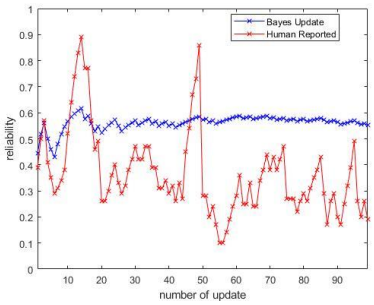

Figure 4. Estimated and Perceived of System reliability of Participant 7

Table 4. Optimal pair of (m, n) for each subject with the largest correlation coefficient $\rho(r^*, r)$, the p value respect to significant level at 0.05, and the RMSE; the corresponding $\rho(r, t)$, $\rho(r^*, t)$, p values and RMSE.

S.	n, m	r^*, r		r, t		r^*, t		S.	n, m	r^*, r		r, t		r^*, t	
		ρ, p	rmse	ρ, p	rmse	ρ, p	rmse			ρ, p	rmse	ρ, p	rmse	ρ, p	rmse
1	7, 50	0.86, .00	0.21	0.13, .65	0.39	-0.09, .37	0.31	40	1, 50	0.93, .00	0.08	0.87, .00	0.09	0.89, .00	0.08
2	20, 10	0.92, .08	0.13	0.92, .08	0.11	1.00, .00	0.01	41	20, 100	0.86, .14	0.04	-0.51, .49	0.07	-0.03, .79	0.07
3	1, 5	0.02, .82	0.39	-0.32, .00	0.46	0.72, .00	0.08	42	3, 20	0.88, .00	0.23	-0.28, .12	0.28	-0.03, .73	0.49
4	20, 100	0.88, .12	0.08	-0.86, .14	0.22	0.00, .98	0.21	43	3, 5	0.51, .00	0.19	0.40, .02	0.30	0.97, .00	0.16
5	4, 20	0.96, .00	0.19	0.73, .00	0.15	0.75, .00	0.24	44	15, 5	1.00, .00	0.12	0.87, .03	0.07	0.80, .00	0.06
6	1, 10	0.76, .00	0.26	0.24, .02	0.50	0.14, .16	0.74	45	2, 100	0.98, .02	0.16	0.96, .04	0.63	0.84, .00	0.50
7	1, 5	0.24, .02	0.25	0.15, .13	0.46	0.86, .00	0.25	46	2, 100	1.00, .00	0.15	0.08, .92	0.24	0.81, .00	0.34
8	1, 5	0.20, .05	0.23	0.13, .20	0.55	0.27, .01	0.33	47	15, 5	0.97, .00	0.15	0.52, .29	0.38	0.26, .01	0.34
9	1, 100	0.94, .00	0.13	0.94, .00	0.13	0.92, .00	0.04	48	20, 20	0.40, .60	0.14	0.23, .77	0.14	0.90, .00	0.12
10	20, 5	0.97, .03	0.21	0.62, .38	0.35	0.45, .00	0.38	49	4, 10	0.86, .00	0.13	0.04, .86	0.52	0.23, .02	0.59
11	2, 5	0.84, .00	0.08	-0.65, .00	0.21	-0.47, .00	0.19	50	1, 50	0.61, .00	0.23	0.75, .00	0.20	0.61, .00	0.05
12	15, 20	0.97, .00	0.06	0.74, .09	0.31	0.43, .00	0.22	51	20, 5	0.78, .22	0.17	-0.80, .20	0.18	0.41, .00	0.18
13	15, 5	0.93, .01	0.18	0.76, .08	0.17	0.65, .00	0.13	52	3, 5	0.62, .00	0.07	0.52, .00	0.19	0.89, .00	0.02
14	2, 5	0.39. .01	0.13	0.80, .00	0.28	0.54, .00	0.20	53	15, 50	0.90, .01	0.20	0.43, .39	0.20	0.73, .00	0.24
15	1, 20	0.95, .00	0.33	0.91, .00	0.45	0.92, .00	0.20	54	10, 100	0.89, .00	0.09	0.89, .00	0.09	1.00, .00	0.01
16	3, 5	0.59, .00	0.08	0.08, .65	0.31	0.51, .00	0.26	55	20, 100	0.99, .01	0.11	0.97, .03	0.25	0.23, .02	0.23
17	1, 5	0.75, .00	0.22	0.34, .00	0.41	0.48, .00	0.23	56	20, 5	0.94, .06	0.13	0.70, .30	0.31	0.41, .00	0.32
18	2, 100	0.93, .00	0.09	0.88, .00	0.35	0.86, .00	0.29	57	20, 50	0.84, .16	0.21	-0.99, .01	0.19	0.12, .22	0.42
19	20, 20	0.94, .06	0.13	0.93, .07	0.39	0.73, .00	0.29	58	20, 20	0.98, .02	0.25	0.92, .08	0.13	0.96, .00	0.07
20	20, 5	0.60. .40	0.11	-0.76, .24	0.27	0.41, .00	0.16	59	20, 20	0.92, .08	0.18	0.92, .08	0.04	0.80, .00	0.12
21	20, 5	0.97, .03	0.08	0.12, .88	0.12	0.80, .00	0.18	60	10, 100	0.17, .66	0.12	0.48, .19	0.04	0.80, .00	0.02
22	1, 5	0.16, .12	0.20	-0.10, .34	0.39	0.79, .00	0.20	61	5, 50	0.82, .00	0.03	0.84, .00	0.07	0.73, .00	0.04
23	20, 50	1.00, .00	0.25	0.61, .39	0.23	0.87, .00	0.08	62	20, 20	1.00, .00	0.08	0.94, .06	0.11	0.72, .00	0.18
24	1,100	0.62, .00	0.18	-0.74, .00	0.32	-0.21, .04	0.18	63	3, 5	0.84, .00	0.17	0.58, .00	0.08	0.72, .00	0.12
25	10, 10	0.79, .01	0.11	0.66, .05	0.40	0.45, .00	0.44	64	15, 10	0.97, .00	0.12	0.77, .07	0.09	0.72, .00	0.09
26	1, 5	0.31, .00	0.31	0.11, .28	0.53	0.60, .00	0.22	65	15, 10	0.98, .00	0.14	0.46, .35	0.10	0.68, .00	0.19
27	15, 5	0.78, .07	0.07	0.43, .39	0.21	0.42, .00	0.20	66	10, 20	0.92, .00	0.10	0.86, .00	0.12	0.87, .00	0.08
28	1, 10	0.42, .00	0.08	0.47, .00	0.07	0.46, .00	0.06	67	20, 20	1.00, .00	0.16	1.00, .00	0.06	1.00, .00	0.00
29	20, 10	0.90, .10	0.31	0.90, .10	0.37	0.84, .00	0.12	68	15, 10	0.95, .00	0.07	0.37, .47	0.14	0.74, .00	0.15
30	20, 10	0.99, .01	0.14	0.49, .51	0.23	0.21, .03	0.27	69	20, 10	0.99, .00	0.07	0.96, .04	0.19	0.96, .00	0.12
31	20, 20	0.85, .15	0.15	0.79, .21	0.13	0.69, .00	0.22	70	7, 100	0.86, .00	0.24	0.12, .67	0.12	0.09, .38	0.10
32	20, 10	0.95, .05	0.15	0.66, .34	0.20	0.64, .00	0.19	71	15, 5	0.87, .02	0.06	0.94, .01	0.06	0.79, .00	0.05
33	10, 20	0.58, .10	0.05	0.57, .11	0.18	0.73, .00	0.22	72	20, 20	0.69, .31	0.10	0.64, .36	0.13	0.77, .00	0.06
34	10, 10	0.96, .00	0.08	0.11, .79	0.17	0.59, .00	0.19	73	4, 5	0.52. .01	0.10	0.40, .05	0.17	0.80, .00	0.08
35	20, 20	0.96, .04	0.17	1.00, .00	0.14	0.81, .00	0.08	74	20, 20	0.89, .11	0.17	-0.79, .21	0.10	0.04, .71	0.11
36	2, 50	0.82, .00	0.16	0.53, .00	0.21	0.69, .00	0.35	75	1, 5	0.02, .88	0.07	0.31, .00	0.23	0.58, .00	0.16
37	20, 10	0.86, .14	0.11	0.87, .13	0.13	0.41, .00	0.17	76	20, 10	0.41, .59	0.24	-0.08, .92	0.24	0.87, .00	0.17
38	10, 5	0.71, .03	0.31	0.21, .59	0.20	0.40, .00	0.26	77	20, 5	1.00, .00	0.10	1.00, .00	0.12	0.93, .00	0.08
39	20, 20	0.97, .03	0.15	0.44, .56	0.14	0.34, .00	0.20	78	15, 5	0.94, .01	0.13	0.71, .11	0.03	0.65, .00	0.10

CONCLUSION

This paper applied Bayesian method to estimate operators' perceived system reliability and predicted the pattern of trust change.

Our results reveal that the process through which human operators estimated automation reliability over repeated interactions can be approximated by Bayesian inference. When the system reliability was higher, the result of Bayesian inference was more accurate. We also noticed that participants could be categorized into three types. Most participants belonged to the first type and their reported system reliability can be accurately estimated by the Bayesian model. For this type of participants, the Bayesian inference can also predict the pattern of trust change. In the future work, computational models can be developed to identify the types of users. Group-specific mathematical models can be proposed to predict the dynamics of trust for each type of participants.

REFERENCES

Baker, C. L., Tenenbaum, J. B., & Saxe, R. R. (2007). Goal inference as inverse planning. In Proceedings of the 29th annual meeting of the cognitive science society.

Barber, B. (1983). Logic and the Limit of Trust. New Brunswick, NJ: Rutgers University Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. Trends in Cognitive Sciences, 10, 294-300.

Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A., & Yanco, H. (2012). Effects of changing reliability on trust of robot systems. The 7th annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12), 73-80.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). The empire of chance. Cambridge: Cambridge University Press.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (2008). Bayesian models of cognition. Cambridge University Press.

Hacking, I. (1975). The emergence of probability. Cambridge: Cambridg University Press.

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. Human Factors, 57(3), 407-434.

Khasawneh, M. T., Bowling, S. R., Jiang, X., Gramopadhye, A. K., & Melloy, B. J. (2003). A model for predicting human trust in automated systems. Proceedings of the 8th Annual International Conference on Industrial Engineering – Theory, Applications and Practice. Origins.

Lee, J. D., & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. Ergonomics, 35(10), 1243-1270.

Lee, J. D., & See, K. A. (2004). Trust in technology: Designing for appropriate reliance. Human Factors, 46(1), 50-80

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. Journal of Cognitive Engineering and Decision Making, 6(1), 57-87.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. Journal of Personality and Social Psychology, 49, 95-112.

Schaefer, K. E.Chen, J. Y. C., Szalma, J. L., Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation. Human Factors, 58(3), 377-400.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. Trends in Cognitive Science, 10, 309-318.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating Effects of User Experience and System Transparency on Trust in Automation. HRI '17 Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, 408-416.

Zuboff, S. (1988). In the Age of the Smart Machine: The Future of Work and Power. New York, NY: Basic Books.

Evaluating Effects of User Experience and System Transparency on Trust in Automation

X. Jessie Yang
University of Michigan
500 S State St
Ann Arbor, MI, USA
xijyang@umich.edu

Vaibhav V. Unhelkar
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA, USA
unhelkar@mit.edu

Kevin Li
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA, USA
kmli@mit.edu

Julie A. Shah
Massachusetts Institute of Technology
77 Massachusetts Avenue
Cambridge, MA, USA
julie_a_shah@csail.mit.edu

ABSTRACT

Existing research assessing human operators' trust in automation and robots has primarily examined trust as a steady-state variable, with little emphasis on the evolution of trust over time. With the goal of addressing this research gap, we present a study exploring the dynamic nature of trust. We defined *trust of entirety* as a measure that accounts for trust across a human's entire interactive experience with automation, and first identified alternatives to quantify it using real-time measurements of trust. Second, we provided a novel model that attempts to explain how *trust of entirety* evolves as a user interacts repeatedly with automation. Lastly, we investigated the effects of automation transparency on momentary changes of trust. Our results indicated that *trust of entirety* is better quantified by the average measure of "area under the trust curve" than the traditional post-experiment trust measure. In addition, we found that *trust of entirety* evolves and eventually stabilizes as an operator repeatedly interacts with a technology. Finally, we observed that a higher level of automation transparency may mitigate the "cry wolf" effect — wherein human operators begin to reject an automated system due to repeated false alarms.

Keywords

Supervisory Control; Trust in Automation; Long-term Interactions; Automation Transparency.

1. INTRODUCTION

The use of robots to assist humans during task performance is growing rapidly. Robots have been deployed for applications such as urban search and rescue (USAR) [1], border patrol [2], forest fire monitoring [3] and military service operations [4, 5], among others. During these tasks, robots are considered an extension of their operators, providing an on-site presence while protecting human users from potential harm [1]. Although teleoperation has been the primary mode of interaction between human operators and remote robots in several applications, increasingly autonomous capabilities including control, navigation, planning and perception [4-7] are being incorporated into robots, with the aim of reducing human operators' workload and stress levels.

One major design challenge for such human-robot partnerships is related to human operators' degree of trust in automated/robotic technology. Trust in automation is defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability [8]". Research has indicated that the calibration between operators' trust and an automated/robotic technology's actual ability is often imperfect [9, 10]. Incidents due to over- or under-trust have been well documented [11, 12]. In order to facilitate proper trust-reliability calibration, extensive amounts of research have been conducted examining the factors influencing human operators' trust in automation [9, 10]. However, most existing research has examined trust as a steady-state variable instead of a time-variant variable, with only few exceptions [13-17].

We, thus, were interested in investigating the dynamic nature of trust. We defined *trust of entirety* as a trust measure that accounts for a human's entire interactive experience with automation. To quantify *trust of entirety*, we investigated a more fundamental research question: is the trust rating reported by a human user at time *t* evaluated on the basis of the user's entire interactive experience with an automated/robotic technology retrospectively from time 0, or of his or her momentary interaction with the technology? Furthermore, we examined how *trust of entirety* evolves and stabilizes over time, and how moment-to-moment trust in automation changes upon automation success or failure.

We conducted a human-subject experiment involving 91 participants performing a simulated military reconnaissance task with the help of an imperfect automation. The participants' subjective trust in automation and behavioral responses were collected and analyzed. Our results indicated that *trust of entirety* is better quantified by the average measure of "area under the trust curve [17]" instead of the traditional post-experiment trust measure. In addition, using a first-order linear time invariant (LTI) dynamical system, we found that *trust of entirety* evolves and eventually stabilizes as a human operator undergoes repeated interactions with automation. Finally, we also observed differences

HRI '17, March 06-09, 2017, Vienna, Austria
© 2017 ACM. ISBN 978-1-4503-4336-7/17/03$15.00
DOI: http://dx.doi.org/10.1145/2909824.3020230

in moment-to-moment trust changes when human users worked with automation of varying degrees of transparency.

2. PRIOR ART AND RESEARCH AIMS

Human trust in automated/robotic technology (henceforth, referred to as trust in automation) is critical to seamless adaptation of technology, and has consequently been of interest to HRI researchers since as early as the 1980s [18]. Issues of trust-reliability mis-calibration continue to be active areas of research related to human-robot teaming in its various forms [12, 19-21]. Existing research, however, has primarily examined trust as a steady-state measure, typically evaluated through questionnaires administered to human operators at the end of their interaction with automation. Assuming that a human interacts with automation for *T* time units during an experiment, we denote this post-experiment measure as *Trust (T)*. In several studies [13-17], researchers have viewed trust as a time-variant measure and elicited human operators' trust in "real time" — i.e., during the interaction. Assuming that this real-time measure of trust is elicited during the interaction at time unit *t (< T)*, we denote it as *Trust (t)*.

Using a simulated pasteurization task, Lee and Moray [13, 14] proposed a time-series model of automation trust. In this task, participants controlled two pumps and one heater, each of which could be set to automatic or manual control. A pump fault was introduced in the task, at which point the pump failed to respond accurately to either manual or automatic control. Based on the simulation, the dynamic variation was analyzed and *Trust (t)* was modeled as a function of *Trust (t-1)* and the automatic control's performance. Similarly, using a memory-recognition task, Yang, Wickens and Holtta-Otto [15] reported moment-to-moment, incremental improvement to trust upon automation success, and moment-to-moment, incremental decline in trust upon automation failure. Moreover, automation failure was found to have a greater influence on trust than success. The results from these studies suggest that human operators' trust calibration is a dynamic process that is sensitive to automation performance.

More recently, several studies examined how the timing of automation failures affects an operator's trust in automation. Sanchez [22] manipulated the distribution of automation failures to be concentrated at either the first or second half of a computer-based simulation task. In this study, participants completed 240 trials of a collision avoidance task in which they were required to avoid hitting obstacles while driving an agricultural vehicle with the aid of an imperfect automation. Trust in automation was reported at the end of the 240 trials, with the results indicating a significant recency effect: participants' trust in automation was significantly lower if automation failures were concentrated in the second half of the experiment.

Desai et al. [23] explored the influence of the timing of robot failures on participants' trust during a robot-controlling task. In their experiment, participants maneuvered a robot along a specified path, searched for victims, avoided obstacles and performed a secondary monitoring task. Participants could drive the robot manually or use an imperfect autonomous driving mode, which was designed to malfunction at the beginning, middle or end of the specified path. Participants reported their degree of trust at the end of experiment; the results indicated that robot failures occurring toward the end of the interaction had a more detrimental effect on trust than failures occurring at the beginning or middle of the interaction. Empirical evidence from both studies [22, 23] supported the detrimental recency effect on post-interaction trust.

In a follow-up study, Desai et al. [17] explored the effect of robot failures on real-time trust. The participants performed the same task as in the prior study, but reported their degree of trust in the robot every 25 seconds. The "area under the trust curve" was used to quantify a participant's trust in automation. Intriguingly, the results from this study showed an opposite trend as compared to the previous studies [22, 23]: robot failures at the beginning of interaction had a more detrimental effect on trust.

Our first objective for the present study is to reconcile these seemingly contradictory findings by answering a more fundamental question: Does the real-time trust rating reported by the users at time *t* account for the entire interaction (beginning at time 0), or only the momentary interaction? We define *trust of entirety* as a trust measure that accounts for one's entire interactive experience with automation, and postulate that if trust at time *t* is evaluated retrospectively, a post-interaction trust rating would be a reliable measure for *trust of entirety*. Alternatively, if trust at time *t* is evaluated on the basis of the momentary interaction, average measure of "area under the trust curve" would be a more appropriate measure.

Second, we examine how *trust of entirety* evolves as a human gains more experience interacting with automation. As discussed earlier, prior research has examined how the timing of automation failures affects trust in automation [17]. Here, we focus on a complementary question: how does trust, specifically *trust of entirety*, evolve as a human undergoes repeated interactions with a system with fixed reliability? During long-term interactions with a robot, while a designer may be unable to control when failures occur, he or she can design for a desired level of reliability. By studying the effect of repeated interactions on trust, we seek to glean insights into estimating human trust in automation over long-term interactions. We posit that a user, upon repeated interactions with a system, eventually achieves a stable value of *trust of entirety.* We denote this final, stable trust value as *Trust (∞)*.

Third, we aim to investigate the effect of automation transparency on moment-to-moment changes to trust (i.e., *Trust (t) – Trust (t-1)*). Automation transparency has been defined as "the quality of an interface pertaining to its ability to afford an operator's comprehension about an intelligent agent's intent, performance, future plans and reasoning process [24]". Previously, Wang, Pynadath, and Hill [19] examined the effect of automatically generated explanations on trust. In their simulation, participants worked with a robot during reconnaissance missions. The robot scanned a city and informed its human teammate of potential danger. Two independent variables were manipulated in this study: robot ability (high- and low-ability conditions) and explanation (low-, confidence-level-, and observation-explanation conditions). The robot scanned eight buildings and made eight decisions per mission. Participants' trust in the robot was measured post-mission. The results indicated a higher degree of trust for high-ability robots and for robots that offered explanations for their decisions. This study shed light on the influence of automation transparency on human operators' trust in automation. Nevertheless, due to the experimental setting, their study did not explore moment-to-moment changes to trust as participants experienced automation successes and failures.

In the present experiment, we manipulated automation transparency through either binary or likelihood alarms. Compared with traditional binary alarms, likelihood alarms provide additional information about the confidence and urgency level of an alerted event [25]. We hypothesize that a high-confidence alert would engender a greater increase in trust upon automation success and a greater decline in trust upon automation failure.

3. METHODOLOGY

We conducted a human-subject experiment to answer the three questions posed in Section 2. Inspired by prior research [4, 5, 26],

Figure 1. Dual-task environment in the simulation testbed. The two images show displays from the simulation testbed for the tracking (top) and detection (bottom) tasks respectively. Participants could access only one of the two displays at a time, and could switch between them.

a military reconnaissance scenario was simulated wherein a human operator supervisory controlled a team of remote robots to gather intelligence with the help of an automated threat detector. Human participants performed 100 repeated interactions with the threat detector. Trust and behavioral responses were collected throughout the experiment. In this section, we detail the experiment setup, design, evaluation and procedure.

3.1 Simulation Testbed

Robots and automation are increasingly being used to support humans during reconnaissance operations. A key function of robots in such applications is to assist humans by gathering information about a remote environment and convey it to the operator. We created an analogue simulation testbed, depicted in Figure 1 that simulates a military reconnaissance scenario.

During the simulation, the human operator was responsible for performing a compensatory tracking task while simultaneously monitoring for potential threats in images of a city provided by a team of four drones. To assist in threat detection, alerts from an automated threat detector were also made available to the human. The participant had the option to trust and thereby accept the decisions of the threat detector as-is, or to personally inspect the images and make his or her own decisions. In this dual-task paradigm, the objective of the human operator was to maximize his or her score, which was a combination of tracking and threat detection performance. We next describe these two tasks in detail.

3.1.1 Tracking Task

A first-order, two-axis compensatory tracking task was programmed based on the PEBL's compensatory tracker task (http://pebl.sourceforge.net/battery.html). Participants using a joystick, moved a green circle to a crosshair located at the center of the display — i.e., minimize the distance between the green circle and the crosshair as shown in Figure 1.

3.1.2 Detection Task

Along with the tracking task, participants were also responsible for monitoring the environment for potential threats.

In each trial, participants received a set of four images from the simulated drones and inspected the presence or absence of threats, with the help of an automated threat detector. We incorporated two types of threat detectors as a between-subject factor, and the reliability of the threat detector was configured according to the signal detection theory (see Section 3.2 for details). An alert was triggered in both visual and auditory modalities.

Participants were asked to report the presence of one or more threats by pushing the "Report" button on the joystick as accurately and as quickly as possible. Along with the detector's alert, the participants had the option of personally inspecting the images. They were allowed to access only one of the two displays — tracking or detection — at a time, and could switch between them using a "Switch" button on the joystick. The participants could perform the tracking task using the joystick throughout the trial, even though they were allowed access to only one display at a time.

During the experiment, participants performed 100 trials of this dual-tasking military reconnaissance mission. Each trial initiated on the tracking display and lasted 10 seconds. The type and performance of the alarm, which varied between the participants, is detailed below.

3.2 Alarm Configuration

We used two types of automated threat detector during the experiment: binary and likelihood. The binary alarm provided one of two alert messages — "Danger" or "Clear" — based on whether it identified the presence of a threat. The likelihood alarm provided a more granular alert: Along with "Danger" or "Clear," it provided two additional alert messages — "Warning" or "Possibly Clear" — implying a lower level of confidence in the detector's decision.

The performance of the automated threat detector was configured based on the framework of signal detection theory (SDT). SDT models the relationship between signals and noises, as well as the threat detector's ability to detect signals among noises [27]. The state of the world is characterized by either "signal present" or "signal absent," which may or may not be identified correctly by the threat detector. The combination of the state of the world and the threat detector's alert results in four possible states: "hit," "miss," "false alarm" and "correct rejection".

Within the context of SDT, two important parameters must be set: the sensitivity (d') of the system when discriminating events rom non-events, and the criterion of the system (c_i) for determining the threshold of an alarm. These parameters are represented in Figure 2. In the present study, the quality of both types of automated threat detector was modeled by manipulating the sensitivity d', which was increased from 0.5 to 3.0 to present an increasing level of automation performance. The first threshold (c_1) was set at 1.0 and was common to both the binary alarm and the likelihood alarm. For the likelihood alarm, along with the first threshold, two additional thresholds were required: c_2, the threshold differentiating dark green ("Clear") and light green ("Possibly Clear") alerts; and c_3, the threshold differentiating red ("Danger") and amber ("Warning") alerts. The values of c_2 and c_3 were set at 0.5 and 3.0, respectively. Benchmarking previous studies [28], the base event rate was set at 30%, indicating that potential threats were present in 30 out of the 100 trials.

3.3 Design

The experiment was carried out according to a repeated-measures, between-subjects design. This design involved two independent variables: alarm type and alarm reliability. The value

Table 1. Four possible states according to SDT

		State of the world	
		Signal	**No Signal**
Threat detector decision	**Signal**	Hit	False alarm
	No Signal	Miss	Correct rejection

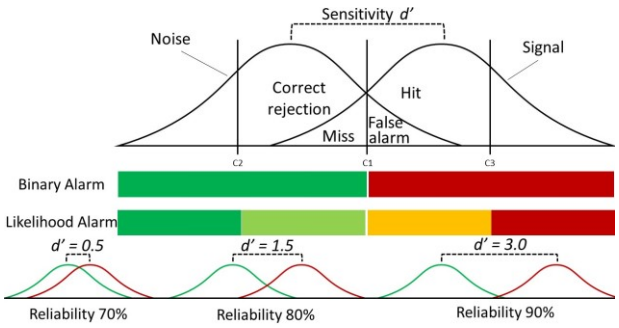

Figure 2. Demonstration of the binary and likelihood alarms, with increasing sensitivity

of alarm reliability was achieved by manipulating alarm sensitivity (*d'*). Two conditions were present for alarm type (binary and likelihood) and three for alarm reliability (70%, 80% and 90%), resulting in six treatment conditions, apart from a control condition. For each of these conditions, based on the associated values of *d'*, *c1*, *c2* and *c3*, the corresponding occurrences of hits, misses, false alarms and correct rejections were computed (Table 2).

A total of 91 participants (average age = 24.3 years, SD = 5.0) with normal or corrected-to-normal vision and without reported color vision deficiency participated in the experiment. They were assigned to one of seven experimental conditions, including six treatment conditions and one control condition. Randomization when assigning experimental conditions was stratified according to participants' self-reported experience playing flight simulation games and first-person shooting games, in order to minimize potential confounding effects.

3.4 Dependent Measures

The dependent variables of interest for the present paper were participants' subjective trust in automation and objective measures of their display-switching behaviors. Working with the same detector, participants completed reconnaissance tasks for 100 sites (100 trials). After each site, participants indicated their subjective trust in the automated threat detector, denoted as *Trust (t)*, using a visual analog scale, with the leftmost anchor indicating "I don't trust the threat detector at all" and the rightmost anchor indicating "I trust the threat detector completely." The visual analog scale was later converted to a 0-100 scale. In addition, for each trial, whether participants switched, and the time at which participants switched their display from tracking to detection were recorded. We used these measures to compute participants' trusting behaviors, which will be discussed in Section 4.1.

3.5 Procedure

Participants signed an informed consent form and provided demographic information. They then received the following description and instructions:

"A group of potential threats has taken over a city, and we are sending you in together with four drones to find out where the threats are before a reinforcement team comes. As a soldier, you have two tasks at the same time: First, you have to make sure that the drones are maintaining level flight. Due to external turbulences, the drones (indicated as the green circle) will be unstable and the green circle will move away from the center (indicated as the crosshair sign). You will control the joystick and move the green circle back to the center as close as possible. At the same time, the four drones will navigate in the city and take surveillance pictures every 10 seconds. The pictures will be sent back to you for threat detection. You need to report to your commander if you identify a potential threat as accurately and as fast as possible by pressing the "Report" button. Due to resource limitations, you can only access one display at a time and you need to press the "Switch" button to switch between the tracking and the detection display. There is an automated threat detector to help you with the task."

If a participant was assigned to a binary alarm condition, they were told the following: "If the detector identifies a threat at a site, the red light in the detector will be on and you will also hear the sound 'Danger.' If the detector identifies there is no threat, the green light will be on and you will hear the sound 'Clear.'

If a participant was assigned to a likelihood alarm condition, they were told the following: "If the detector identifies a threat at a site, either the red light or the amber light will be on, and you will also hear the sound 'Danger' or 'Warning,' respectively. The red light and the 'Danger' sound indicate a higher level of confidence and the amber light and 'Warning' sound indicates a lower level of confidence. If the detector identifies no threat, either the dark green or the light green light will be on, and you will hear the sound 'Clear' or 'Possibly Clear.' The dark green light and the 'Clear' sound indicate a higher level of confidence that the site is safe. The light green light and the 'Possibly clear' sound mean a lower level of confidence that the site is safe.''

Table 2. Alarm configurations and corresponding numbers of hits, misses, false alarms and correct rejections

Reliability = 70%					
Binary alarm			Likelihood alarm		
Alert	Threat	Clear	Alert	Threat	Clear
Danger	9	11	Danger	5	5
			Warning	4	6
Clear	21	59	Possibly clear	6	11
			Clear	15	48
Reliability = 80%					
Binary alarm			Likelihood alarm		
Alert	Threat	Clear	Alert	Threat	Clear
Danger	21	11	Danger	15	5
			Warning	6	6
Clear	9	59	Possibly clear	4	11
			Clear	5	48
Reliability = 90%					
Binary alarm			Likelihood alarm		
Alert	Threat	Clear	Alert	Threat	Clear
Danger	29	11	Danger	28	5
			Warning	1	6
Clear	1	59	Possibly clear	1	11
			Clear	0	48

After the introduction, participants completed a practice session consisting of a 30-trial block of the tracking task only, followed by an eight-trial block including both the tracking task and the detection task. Hits, misses, false alarms and correct rejections were illustrated during the eight practice trials of combined tasks. The participants were told that the alerts from the automated threat detector may or may not be correct. The subsequent experimental block consisted of 100 trials, lasting approximately 60 minutes with a 5-minute break at the halfway point. Each participant received compensation consisting of a $10 base plus a bonus up to $5. The compensation scheme was determined through a pilot study, incentivizing participants to perform well on both tasks.

4. ANALYSIS AND RESULTS

In this section, we discuss the observations, data and results from our experiment. We first examine which user-reported trust measures are good indicators of *trust of entirety*. Next, we present a novel model to explain the evolution of *trust of entirety* over time. Finally, we present our findings on the relationship between automation transparency and moment-to-moment trust changes. Data from participants in the control group was excluded from the subsequent analysis, as they did not receive any automation aid and did not report subjective trust.

4.1 Indicators of Trust of Entirety

In prior literature, two measures of trust have been used to quantify *trust of entirety*: $Trust_{end}$, the trust rating elicited after the terminal trial T, and $Trust_{AUTC}$, the area under the trust curve. For our experiment, we computed these quantities as follows (note that the computation of $Trust_{AUTC}$ included averaging of trust across number of interactions):

$$Trust_{end} = Trust_T$$

$$Trust_{AUTC} = \frac{1}{T}\sum_1^T Trust_t, \text{ where T = number of interactions}$$

To examine whether $Trust_{end}$ or $Trust_{AUTC}$ corresponds to *trust of entirety* more appropriately, we calculated the correlation between the two subjective trust measures and the participants' trusting behaviors including their reliance and compliance behaviors. Reliance has been defined in prior literature as the human operator's cognitive state when automation indicates no signal (no threat); compliance represents the human operator's cognitive state when automation indicates a signal (threat) [29].

In the present study, we measured both response rate (RR) and response time (RT). Reliance is characterized by trusting the automation to indicate "Clear" or "Possibly Clear" in the absence of a threat, and thus no switch or a slower switch from the tracking task to the detection task. Compliance is characterized by trusting the automation to signal "Danger" or "Warning" in the presence of one or more threats, and thus reporting threats blindly with no switch to the detection task, or a rapid switch from the tracking task to the detection task. Further, we calculated the difference between reliance RT and compliance RT. This measure eliminates potential confounding effects due to participants' intrinsic characteristics of switching behaviors (i.e., participants may switch more quickly or slowly regardless of the alerts [22]).

$Compliance_{RR}(C_{RR}) =$
$Prob(report\ without\ switch\ |danger\ or\ warning\ alerts)$

$Reliance_{RR}(R_{RR}) =$
$Prob(not\ report\ without\ switch\ |\ clear\ or\ possibly\ clear\ alerts)$

$Compliance_{RT}(C_{RT}) =$
$Time\ (first\ switch\ |danger\ or\ caution\ alerts)$

$Reliance_{RT}(R_{RT}) =$
$Time\ (first\ switch|clear\ or\ possibly\ clear\ alerts)$

Table 3. Pearson correlation coefficient between trust measures and participants' trusting behavior ($^{*}p < .05$; $^{**}p < .01$)

	C_{RR}	C_{RT}	R_{RR}	R_{RT}	R_{RT} - C_{RT}
$Trust_{end}$	n.s.	n.s.	n.s.	n.s.	n.s.
$Trust_{AUTC}$	.31**	n.s.	.31**	n.s.	.26*

Table 3 summarizes results from Pearson's correlation analysis. $Trust_{AUTC}$ was significantly correlated with C_{RR}, R_{RR}, and (R_{RT} - C_{RT}), while $Trust_{end}$ was not significantly correlated with any of the behavioral measures. These results indicate that a user's degree of trust reported at time t is more influenced by their momentary interaction with automation. Therefore, we claim that $Trust_{AUTC}$ is a more appropriate measure of *trust of entirety*.

This finding could explain the seemingly contradictory findings in previous studies [17, 22, 23]: when automation failures occurred toward the end of an experiment, it resulted in a momentary decline in trust. As $Trust_{end}$ was used in these studies to quantify participants' entire interactive experience with automation, it was more severely affected as compared with a condition under which automation failures occurred at the beginning of the interactive process.

For clarification, from this point onward we use $Trust_e$ to denote *trust of entirety*. Further, in subsequent analysis $Trust_{AUTC}$ is used as an indicator for trust of entirety, $Trust_e$.

4.2 Effect of Experience on $Trust_e$-Reliability Calibration

Issues due to over- and under-trust have been a challenge for the adoption of automation, highlighting the need to investigate not only trust-reliability calibration but also how it evolves with experience. Using repeated measurements of self-reported trust, we assessed how $Trust_e$ -reliability calibration varied across trials.

A trust-reliability calibration curve depicts the correspondence between trust and reliability over a wide spectrum of automation reliability [11]. To plot the $Trust_e$-reliability calibration curve, $Trust_e$ was regressed against automation reliability. Figure 3 depicts the slope variation for the $Trust_e$-reliability calibration curve with respect to automation experience (trial number) and Figures 4 and 5 show the calibration curves for the 1st, 50th, and 100th trial, respectively, with the black line indicating the regression line, the blue lines the 95% confidence interval and the red lines the 95% predictive band. Figures 3-5 indicate that the calibration curves change with automation experience for both alarm types. In addition, for both alarm types, the calibration curve became steeper as human operators gained more experience with the threat detector. Further, the slope of the likelihood alarm curve increased more rapidly than that of the binary alarm curve.

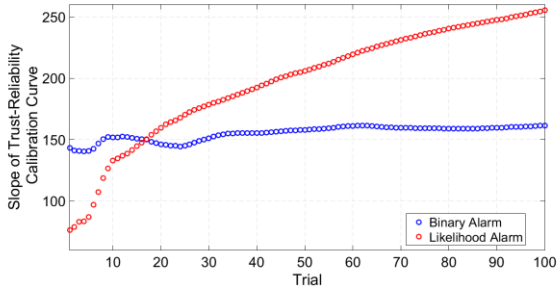

Figure 3. Variation of the slope of the $Trust_e$-calibration curve with automation experience

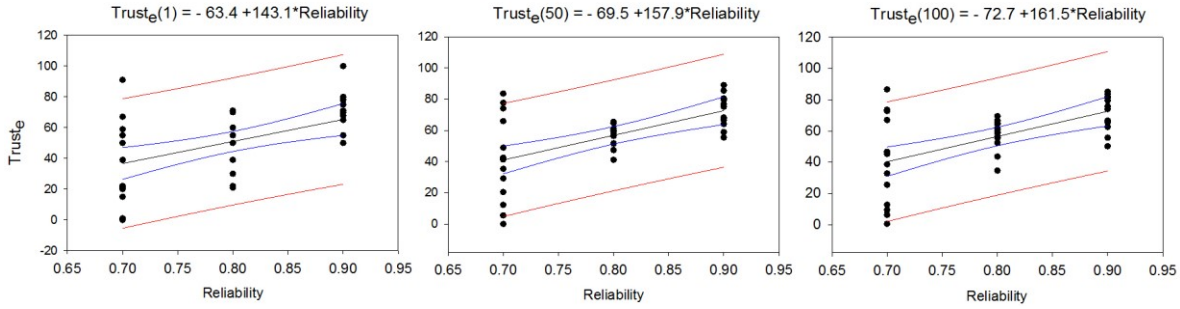

Figure 4. Variation of *Trust*$_e$-reliability calibration with automation experience for binary alarm

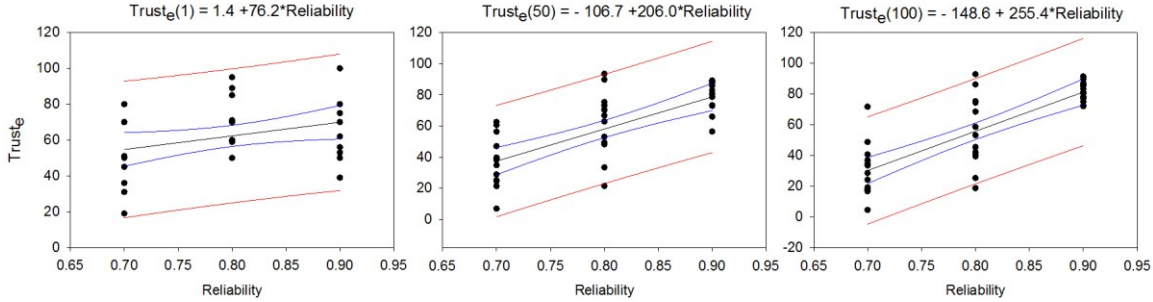

Figure 5. Variation of *Trust*$_e$-reliability calibration with automation experience for likelihood alarm

4.3 Effect of Experience on Trust$_e$

To further understand how trust evolves with experience, we analyzed *Trust*$_e$ with respect to automation experience (trial number) for each of the six treatment conditions. We found that users' trust in automation evolved over time, and that change in trust, averaged across all users, exhibited an asymptotic stabilizing trend for each condition. To explain this trend, we propose a model for the evolution of trust over time. This model is inspired by the theory of dynamical systems, which has previously been used for modeling cognitive processes such as the forgetting curve [30].

We hypothesized that for a robot or automation system that does not involve a learning or adaptive component — i.e., a system with fixed performance/reliability over time — the trust of the *average user* converges to a value *Trust (∞)* as he or she gains experience with the system. Further, the evolution of trust over time can be modeled as the response of a first-order linear time invariant (LTI) dynamical system to a constant (step) input signal.

The proposed model is described mathematically as follows:

$$\frac{d\,Trust_e}{dt} + \frac{Trust_e}{\tau} = G\,u(t)$$

t corresponds to time (experience with automation), τ corresponds to the 'time constant' of the system, G corresponds to the system gain and $u(t)$ corresponds to unit step input. In the context of our experiment, these quantities further relate as follows: t represents the trial number; τ represents a quantity proportional to the number of trials needed for trust to reach its final, stable value; and the constant (step) input corresponds to a system with fixed reliability.

Upon solving the above first-order differential equation, the evolution of trust with experience can be represented as follows:

$$w = \exp(-t/\tau)$$

$$Trust_e(t) = Trust_e(final) * [1 - w] + Trust_e(initial) * [w]$$

$$Trust_e\ Change\ (t) = Trust_e\ Change(final)[1 - \mathrm{w}]$$

We fitted the above equation to the trust measurements recorded during our experiment. The data was fit to the mean value of trust for different instances of automation — i.e., different reliability condition for each alarm type. We used Matlab's nonlinear least squares method for curve fitting, and estimated the initial trust using the mean trust level during the first interaction. The resulting plots are depicted in Figures 6 and 7 for the binary and likelihood alarm, respectively. The plots include a scatter plot of the data and the fitted curve along with its 95% confidence interval. Goodnesss of fit for each curve is quantified using "adjusted R-squared" and is listed in Table 4, which also includes the estimated value of the time constant and the estimated asymptotic value of trust.

The adjusted R-squared values indicate that the proposed first-order dynamical systems is a good fit for the empirically observed data. The goodness of fit is higher for the likelihood alarm. Further, the estimated final values of trust as determined by our first-order model vary proportionally with system reliability. Additionally, we observed two interesting patterns. First, the time constant is greater for the likelihood alarm; this implies that users require more interaction with in order to arrive at a stable trust value for the likelihood alarm. This may be due to the greater number of alternatives associated with the likelihood alarm compared with the

Table 4. Results from the first-order model of Trust

Alarm type	Reliability	Adjusted R-squared	Estimated Time Constant	Estimated *Trust(∞)*
Binary	70%	0.662	11.48	40.81
	80%	0.963	19.23	57.24
	90%	0.896	20.50	72.83
Likeli-hood	70%	0.962	35.77	30.69
	80%	0.994	49.68	53.48
	90%	0.995	41.72	83.15

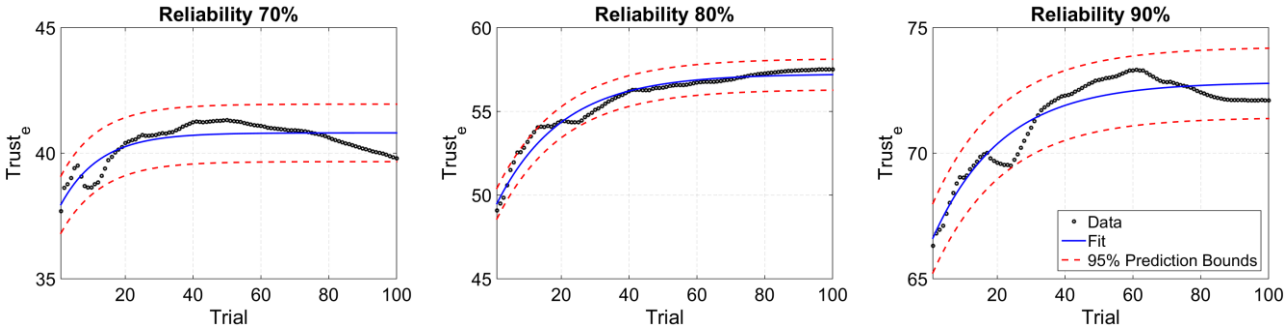

Figure 6: Variation of *Trust*$_e$ (averaged across participants) with automation experience for binary alarm.

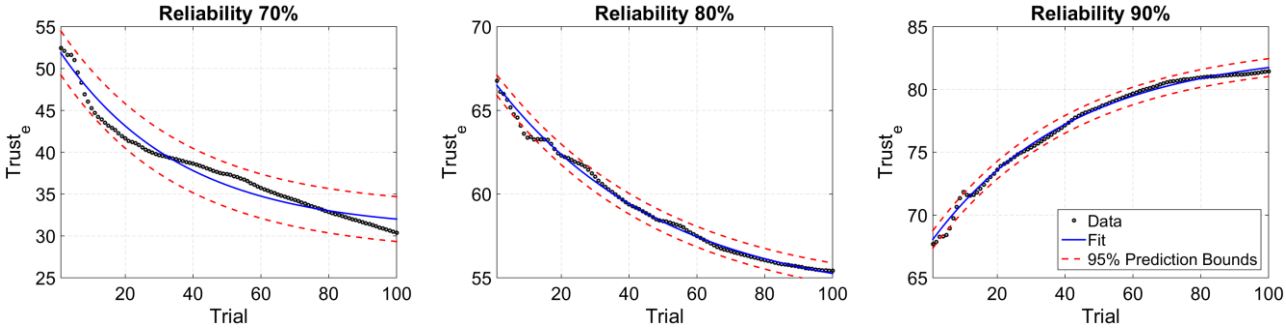

Figure 7. Variation of *Trust*$_e$ (averaged across participants) with automation experience for likelihood alarm. Notice the change in the magnitude of *Trust*$_e$ (y-axis) across the three plots for both the binary (Fig. 6) and likelihood (Fig. 7) alarm. We observe that higher automation reliability results in higher value of trust.

binary alarm, resulting in users requiring more time to create a stable, mental model of the likelihood alarm.

Second, we observed that while using binary alarms, participants' *Trust*$_e$ increased with repeated interactions with automation for all three automation reliability levels, whereas when using likelihood alarms, *Trust*$_e$ decreased over time at reliability of 70% and 80% and increased at reliability of 90%. This variation in *Trust*$_e$ evolution patterns may be explained by the interplay between operators' initial expectation of automation and their subsequent observation of automation's performance [9]. Studies have shown that people have higher initial expectation and trust when automation is portrayed as an "expert" system [31, 32]. Likelihood alarms may be perceived more "intelligent" compared to binary alarms of the same reliability and engender higher initial trust. As participants interacted with the threat detector, they adjusted their trust to reflect automation's true performance. Trust decrement may reflect participants' initial over-expectation and subsequent decrement of trust, whereas trust increment reflects initial under-expectation and subsequent increment of trust.

Note that caution is warranted when interpreting the estimates of the LTI model described above. The model is obtained using the average measurements of trust across participants; thus, it allows for estimation of the degree of trust likely to be exhibited by the average user. For instance, the final value of trust obtained by the empirical fit provides the average degree of trust that might be observed across users. Although the model does not allow for predictions regarding the evolution of trust for a single user, its utility lies in estimating the average value of trust in a system across multiple users. Further, we claim the applicability of this model only for systems with fixed reliability/performance; this may or may not extend to systems that adapt or learn during interaction.

4.4 Effect of automation transparency on momentary trust change

In order to examine the effect of automation transparency on the change of moment-to-moment trust, $Trust_t - Trust_{t-1}$, we conducted the following tests: (i) paired sample t-tests to compare the differences between high- and low-likelihood alerts, (ii) independent sample t-tests to compare the differences between binary and high-likelihood alerts, and between binary and low-likelihood alerts. Note that paired-sample t-tests have greater statistical power than independent sample t-tests.

Figures 8-11 depict momentary change of $Trust_t$ for hits, false alarms, correct rejections and misses. When the threat detector's decisions were hits, there was a marginally significant difference between high- and low-likelihood alerts (1.22 vs. 0.71, paired sample $t(38) = 1.946$, $p = .06$), indicating that a correct alert of threat presence with high confidence led to a greater increase to $Trust_t$ in comparison with $Trust_{t-1}$. When the threat detector gave false alarms, $Trust_t$ decreased. Further, the comparisons indicated a significant difference between high- and low-likelihood alerts (-4.95 vs -1.49, paired sample $t(38) = -3.11$, $p < .01$) and between binary alerts and low-likelihood alerts (-3.31 vs -1.49, independent $t(76) = -2.37$, $p = .02$).

$Trust_t$ increased when the threat detector correctly identified the absence of threats. Moreover, we observed a significantly greater improvement to $Trust_t$ for high-likelihood alerts compared with low-likelihood alerts (0.72 vs 0.40, paired sample $t(38) = 2.085$, $p = .04$), and a marginally significantly greater improvement for binary alerts compared with low-likelihood alerts (0.76 vs 0.40, independent $t(76) = 1.848$, $p = .07$). When the threat detector missed potential threat, there was a decrease in $Trust_t$, with no statistically significant differences between high- and low-likelihood alerts, between binary and high-likelihood alerts or between binary and low-likelihood alerts.

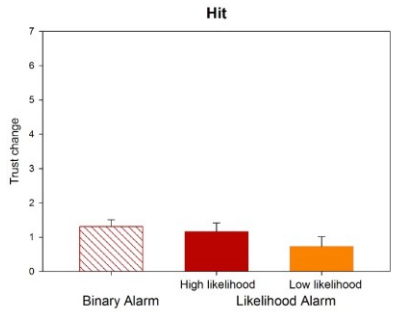

Figure 8. Momentary change of *Trust$_t$* for hits

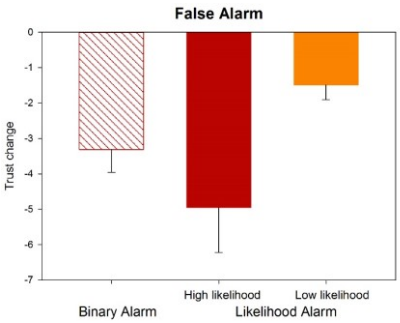

Figure 9. Momentary change of *Trust$_t$* for false alarms

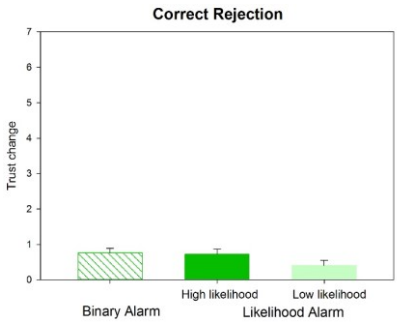

Figure 10. Momentary change of *Trust$_t$* for correct rejections

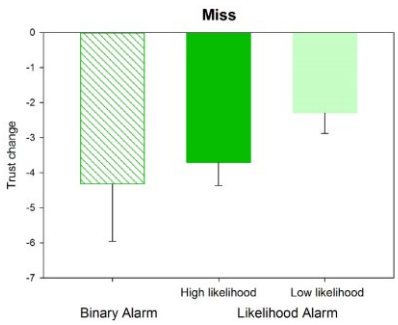

Figure 11. Momentary change of *Trust$_t$* for misses

5. DISCUSSIONS

Our first objective was to determine whether a human operator's trust rating at time *t* is evaluated on the basis of his or her entire interactive experience or according to the momentary interaction with automation. Our results indicated that trust at time *t* was evaluated according to the momentary interaction, and that *trust of entirety* was better quantified by $Trust_{AUTC}$ compared with $Trust_{end}$. This finding has important implications for trust measurement during human-automation/robot interaction — specifically, merely administering a trust survey at the end of an experiment is inadequate if the intent is to measure human participants' degree of trust in an automated/robotic technology over the course of the entire interactive process. Continuous trust measure in real time is necessary to achieve this goal.

The second objective of the present study was to explore how *trust of entirety* evolves as a human gains more experience interacting with automation. Our proposed first-order LTI model suggested that *trust of entirety* evolved and stabilized as an operator interacted more with the automated system. Interestingly, we observed a larger time constant for the likelihood alarm, suggesting that human operators require longer interaction with this type of alarm in order to arrive at a stable value of trust. This finding is potentially attributable to the high- and low-likelihood information associated with the alarm, which may require additional trust calibration before a steady state is reached. Additionally, we observed variations in patterns of trust evolution, which could be explained by the interplay between human participants' initial expectation of automation and their subsequent adjustment of trust in automation.

Our third objective was to investigate the influence of automation transparency on human operators' moment-to-moment trust changes. Increasing automation transparency has been proposed as a method of increasing a human operator's trust in automation [22]. Findings from this study confirm that high-likelihood alerts engender a greater increase to momentary trust upon automation success, as well as a greater decline in momentary trust upon automation failure. Our results also shed light upon the underlying reason for the benefits of increasing automation transparency: higher automation transparency may mitigate the "cry wolf" effect.

The "cry wolf" effect is a phenomenon commonly observed in high-risk industries in which the threshold to trigger an alarm is often set very low in order to capture every critical event [9]. This low threshold, however, inevitably results in false alarms, which can cause human operators to question or even abandon the automated technology. The significant difference we observed in the response to low-likelihood and binary alerts suggests that human participants were still able to retain their trust in automation if the false alarm was provided through low-likelihood alerts. It is possible that users are less inclined to interpret these false alarms as false since the low-likelihood alerts merely suggest that a threat may exist, rather than explicitly confirm the presence of a threat.

6. CONCLUSION

Existing research examining human trust in automation and robots has primarily examined trust as a steady-state variable, with little emphasis on the evolution of trust over time. The present study explored the dynamic nature of trust.

We defined *trust of entirety* as a trust measure that accounts for a human's entire interactive experience with automation. Using a simulated reconnaissance task, we conducted a human-subject experiment (N=91) and found that $Trust_{AUTC}$ is a more appropriate measure for *trust of entirety*. The present study also showed that *trust of entirety* evolves and stabilizes over time, and demonstrated that a higher level of automation transparency may mitigate the "cry wolf" effect.

7. ACKNOWLEDGEMENTS

This work is supported by the SUTD-MIT postdoctoral fellows program.

8. REFERENCES

[1] Murphy, R. R. 2004. Human-robot interaction in rescue robotics. *IEEE Transactions on Systems, Man, and Cybernetics,* 34, 2, 138-153.

[2] Girard, A. R., Howell A. S., and Hedrick, J. K. 2004. Border patrol and surveillance missions using multiple unmanned air vehicles. *The 43rd IEEE Conference on Decision and Control,* 620-625.

[3] Casbeer, D. W., Kingston, D. B., Beard, R. W. and McLain, T. W. 2006. Cooperative forest fire surveillance using a team of small unmanned air vehicles. *International Journal of Systems Science*, 37, 6, 351-360.

[4] Chen, J. Y. C. 2010. Robotics operator performance in a multi-tasking environment: Human-Robot Interactions in Future Military Operations. Ashgate Publishing, 294-314.

[5] Chen, J. Y. C. and Barnes, M. J. 2008. Robotics operator performance in a military multi-tasking environment. *The 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI '08)*, 279-286.

[6] Talamadupula, K., Briggs, G. Chakraborti, T. Scheutz M. and Kambhampati, S. 2014. Coordination in human-robot teams using mental modeling and plan recognition, *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2957-2962.

[7] Pratt, G. and Manzo, J. 2013. The DARPA Robotics Challenge [Competition]. *IEEE Robotics & Automation Magazine*, 20, 2, 10-12.

[8] Lee, J. D. and See, K. A. 2004. Trust in technology: Designing for appropriate reliance. *Human Factors*, 46, 1, 50-80.

[9] Hoff, K. A. and Bashir, M. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors*, 57, 3, 407-434.

[10] Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J. and Parasuraman, R. 2016. A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors*, 53, 5, 517-527.

[11] Wickens, C. D., Hollands, J. G., Banbury, S. and Parasuraman, R. 2013. *Engineering Psychology & Human Performance*. Pearson Education.

[12] Robinette, P., Li, W., Allen, R., Howard, A. M. and Wagner, A. R. 2016. Overtrust of Robots in Emergency Evacuation Scenarios. *The 11th ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*, 101-108.

[13] Lee, J. D. and Moray, N. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35, 10, 1243-1270.

[14] Lee, J. D. and Moray, N. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human Computer Studies*, 40, 1, 153-184.

[15] Yang, X. J., Wickens, C. D. and Hölttä-Otto, K. 2016. How users adjust trust in automation: Contrast effect and hindsight bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 60, 1, 196-200.

[16] Manzey, D., Reichenbach, J. and Onnasch, L. 2012. Human Performance Consequences of Automated Decision Aids: The Impact of Degree of Automation and System Experience. *Journal of Cognitive Engineering and Decision Making*, 6, 1, 57-87.

[17] Desai, M., Kaniarasu, P., Medvedev, M., Steinfeld, A. and Yanco, H. 2013. Impact of robot failures and feedback on real-time trust. *The 8th ACM/IEEE international conference on Human-robot interaction (HRI '13)*, 251-258.

[18] Muir, B. M. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, 27, 5, 527-539.

[19] Wang, N., Pynadath, D. V. and Hill, S. G. 2016. Trust Calibration within a Human-Robot Team: Comparing Automatically Generated Explanations. *The 11th ACM/IEEE International Conference on Human Robot Interaction (HRI '16),* 109-116.

[20] Lohani, M., Stokes, C., McCoy, M., Bailey, C. A. and Rivers, S. E. 2016. Social Interaction Moderates Human-Robot Trust-Reliance Relationship and Improves Stress Coping. *The 11th ACM/IEEE International Conference on Human Robot Interaction (HRI '16)*, 471-472.

[21] Bartlett, C. E. and Cooke, N. J. 2015. Human-Robot Teaming in Urban Search and Rescue. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59, 1, 250-254.

[22] Sanchez, J. 2006. *Factors that affect trust and reliance on an automated aid*. Georgia Insititue of Technology.

[23] Desai, M., Medvedev, M., Vázquez, M., McSheehy, S., Gadea-Omelchenko, S., Bruggeman, C., Steinfeld, A. and Yanco, H. 2012. Effects of changing reliability on trust of robot systems. *The 7th annual ACM/IEEE international conference on Human-Robot Interaction (HRI '12)*, 73-80.

[24] Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D. and Procci, K. 2016. Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management. *Human Factors*, 58, 3, 401-415.

[25] Sorkin, R., Kantowitz, B. H. and Kantowitz, S. C. Likelihood alarm displays. 1988 *Human Factors*, 30, 4, 445-459.

[26] Wickens, C. D., Levinthal, B. and Rice, S. 2010. Imperfect reliability in ummaned air vehicle supervison and control: Human-Robot Interactions in Future Military Operations. Ashgate Publishing, 193-210

[27] Tanner, W. P. J. and Swets, J. A. 1954. A decision-making theory of visual detection. *Psychological Review*, 61, 6, 401-409.

[28] Wiczorek, R. and Manzey, D. 2014 Supporting attention allocation in multitask environments: Effects of likelihood alarm systems on trust, behavior and performance. *Human Factors*, 56, 7, 1209-1221.

[29] Dixon, S., Wickens, C. D. and McCarley, J. M. 2007 On the independence of reliance and compliance: Are false alarms worse than misses? *Human Factors*, 49, 4, 564-572.

[30] Ebbinghaus, H. 1913. *Memory: A Contribution to Experimental Psychology*. Columbia University, New York City.

[31] Madhavan, P., & Wiegmann, D. A. 2007. Effects of information source, pedigree, and reliability on operator interaction with decision support systems. *Human Factors*, 49, 5, 773–785

[32] de Vries, P., & Midden, C. 2008. Effect of indirect information on system trust and control allocation. *Behaviour & Information Technology,* 27, 1, 17–29.