# A PROCESS MODEL OF TRUST IN AUTOMATION: A SIGNAL DETECTION THEORY BASED APPROACH

Jorge Zuniga[1], Malcolm McCurry[2], J. Gregory Trafton[3]

George Mason University[1] Fairfax, VA

Exelis, Inc[2] Alexandria, VA

Naval Research Laboratory[3] Washington, DC

This paper discusses the first experiment in a series designed to systematically understand the different characteristics of an automated system that lead to trust in automation. We also discuss a simple process model, which helps us understand the results. Our experimental paradigm suggests that participants are agnostic to the automation's behavior; instead, they merely focus on alarm rate. A process model suggests this is the result of a simple reward structure and a non-explicit cost of trusting the automation.

## INTRODUCTION

Trust in automation is important because it guides understanding of user interactions with systems. Trust has been linked with user reliance on automation; it has also been connected with different types of errors such as misuse, disuse, and abuse (Parasuraman & Riley, 1997). Discussions of trust in automation inevitably involve its performance in the environment (Lee & See, 2004; Parasuraman & Riley, 1997; Wickens & Dixon, 2007; Yeh & Wickens, 2001). This performance is generally communicated in terms of correctness (Muir & Moray, 1996; Parasuraman & Miller, 2004; Wiegmann, Rich, & Zhang, 2001), but it is partly dependent on the exact types of behaviors that the automation exhibits (Dixon, Wickens, & McCarley, 2007; Meyer, 2001), such as the types of errors it makes.

In fact, automation performance can be characterized in terms of signal detection theory (SDT; Green & Swets, 1966). From this perspective, correct behaviors by the system are analogous to hits and correct rejections, whereas errors can be identified as misses or false alarms (FA). For example, computer-aided diagnosis (CAD) helps radiologists identify tumors or other diseases in the radiology industry. But what happens when CAD fails to identify a tumor (miss), or tells the radiologists that a tumor is present when it is not (false alarm)? The cost of failing to identify cancer early is often times lethal; while prescribing unneeded treatment is costly and often times also dangerous. The dangers of misuse, disuse, or abuse (Parasuraman & Riley, 1997) of automation are clear. Paramount to understanding what drives these sources of error is trust calibration.

Dixon et al. (2007) explored differences in user behaviors closely tied with trust calibration. They compared no automation, perfect automation, and two types of error-prone systems. One was miss-prone which had a 20% hit rate, but exhibited perfect FA behavior (0% FA). The other system was FA-prone automation, that system made 80% FA, but exhibited perfect hits (100% hits). They measured reliance (in our example: agreeing when CAD identifies a tumor), as well as, compliance (in our example: agreeing with CAD when it suggests there is *no* tumor). Dixon et al. (2007) found that FA-prone automation negatively affects both reliance and compliance while miss-prone automation only affected compliance.

While we also manipulated misses and false alarms in our research, we were more interested in how users would change their behaviors when faced with equally imperfect types of automation. Are participants more attuned to misses than FA? What dictates whether a user will start paying attention to FA over hits? SDT allows for the calculation of sensitivity (d'). Sensitivity is a measure of the combined rate of hits and FAs of a system; it communicates the accuracy of the system in identifying the signal from the noise. However, it is possible to create equally sensitive systems with vastly different behavior patterns.

Table 1 Each system presented here represents a condition in our experiment.

We designed four different automated systems. All four systems were equally sensitive (d' = 2.32), and ranged from low hit rate/low false alarm rate, to high hit rate/high false alarm rate (see Table 1). This approach allows us to identify what characteristic of the system users are attuned to by analyzing the effects of each type of equally sensitive system on user's patterns of responses to the automation (Table 2).

Table 1

*Breakdown of Automated systems*

| System | True Positive Rate | False Positive Rate |
|---|---|---|
| 91/15 | 97% | 15% |
| 85/10 | 85% | 10% |
| 75/5 | 75% | 5% |
| 67/3 | 67% | 3% |

## Cognitive Model

In addition to conducting an experiment we also built a process model of the task using the ACT-R (Adaptive Control of Thought-Rational) cognitive modeling architecture (Anderson, 2007). ACT-R is a theoretically grounded cognitive architecture that allows researchers to create process models that are able to mimic human cognition. ACT-R allows researchers to model the internal process which are being used by a user as they perform a task. This approach has been used in the past to better understand other cognitive processes such as errors, vigilance, and driving (Gunzelmann, Byrne, Gluck, & Moore Jr, 2009; Salvucci, 2006; Trafton, Altmann, & Ratwani, 2011). ACT-R is divided into several modules, which can be equated to different parts of information processing theory. For this task we primarily took advantage of one of the learning components of ACT-R (utility learning) which is based on the difference learning equation (Fu & Anderson, 2006). It is also very similar to the Rescorla-Wagner learning rule (Rescorla & Wagner, 1972).

# METHODS

## Participants

Sixty George Mason University undergraduate students participated in this study. They received course credit for their participation.

## Task and Materials

Participants were told that they were interacting with a simulated mining environment. They engaged in a dual-task paradigm in which they had to operate a drill and send the minerals they collected to a warehouse by monitoring and responding to the appropriate color of a cart in a secondary hidden window. They were assisted by an automated cueing system. The main task consisted of tracking a moving box with the mouse as it traveled around

Table 2

*Hypothesis Table*

| | If Participants are sensitive to… | | | | |
|---|---|---|---|---|---|
| | Hits | FA | Misses | d' | Alarms |
| Cued Switch (CS) | More CS with higher hits | Less CS with higher FA | Less CS with lower hits | No pattern | More CS with higher Alarm rates |
| Uncued Switch (US) | Less US higher hits | More CS with higher FA | More US with lower hits | No pattern | Less US with lower Alarm rates |
| Reaction Time to Cue | Faster with higher hits | Slower with higher FA | Slower with higher hits | No pattern | No Pattern |
| Ignored Cue (IC) | Less IC with higher hits | Increased IC with higher FA | No Pattern | No pattern | No Pattern |

Table 2 Cued switches represent checking the cart whenever the automation suggested. Uncued switches are when users switched without being prompted by the automation. Ignored cues are when the automation suggested switching but users did not do so. Finally, reaction time is the time it took the participant to switch after the automation suggested that they switch

the screen (the drill), while having to monitor a secondary hidden window for a changing color box (the cart) as a secondary task. Participants switched windows by clicking on any one of four buttons located on the corners of the screen. The buttons switched back and forth between both windows.

The goal of the task was to maximize the amount of minerals collected. Keeping the mouse inside the moving box accrued minerals at a rate of 3 minerals per second. Additionally, participants had the opportunity of earning 100 extra minerals by responding, using the spacebar, whenever the box in the secondary window turned red; however, if they pressed the spacebar when it was blue they lost 50 minerals. Participants had to switch to the cart view before making a response. The cost of incorrect response was set up in order to ensure participants actually looked at the cart before responding.

An automated system alerted participants of a cart that was ready by chiming an audible tone. Participants were instructed that the tone was indicative of the automated system sensing the cart was ready to go. However, this automated system was not perfect in that while keeping d' constant at 2.32, we manipulated the exhibited behavior of the automated system as shown in Table 1. For example, in the 91/15 condition the automated system was accurate in sounding the cue to a full cart 91% of the time (hit), however, 15% of the time that the cart was not full it also presented the cue (FA).

The task ran on a Dell laptop (Intel i7-3520 @ 2.90 MHz, 4GB RAM, Win7 32bit) with a Dell P2210 22" monitor at 1680 x 1050 resolution.

### Design and Procedure

This was a between subjects design. Participants were first told their goal in the task (to maximize minerals) and then exposed to the interface through screenshots. They were then introduced to the automated system and the possible behaviors it could exhibit (hits, misses and FA) through a brief 3 trial introductory session. All participants first experienced a hit, then a false alarm, and finally a miss. After this brief introduction, participants engaged in a 3 minute training session. All participants interacted with an 80% hit rate and 30% false alarm rate automation during the training. Participants then began the main task and the experimenter exited the room. After the main session was over, participants had the opportunity to provide comments, after which they were debriefed and thanked for participating.

We measured how many times participants exhibited each of 3 different types of behaviors. *Cued* switches are times in which participants switched after an alarm had sounded. *Uncued* switches are any times that participants switched to the cart without any alarm from the automated system. *Ignored Cues* were any time that the alarm was sounded by the automation but the participant did not respond. Finally, *Reaction Time* was also measured as the time between the automation alarm and the time when the participant clicked the button to switch, it was only calculated for cued switches.

## RESULTS AND DISCUSSION

For ease of understanding we will discuss conditions in terms of their hit rate and false alarm rate, e.g. the condition with 91% hits and 15% false alarms will be referred to as condition "91/15". We compared mean Cued Switch behavior over the different condition using a one way ANOVA. There was a main effect for condition, $F(3, 56) = 17.98$, $MSE = 203$, $p < .0.001$, $\eta^2 = .49$. Tukey's HSD test shows significant differences between all the conditions except for between condition 85/10 - 91/15, and 67/3 - 75/5. As can be seen in Figure 1 there was an overall trend of increasing cued switches with increasing alarm (Hit + FA) rates. Had the participants been impacted by the increasing number of FA, we would see a decreasing trend of switching to the cue. However, this trend does provide some support for participants being impacted by hits just overall alarm rate, which we discuss further along in the paper.

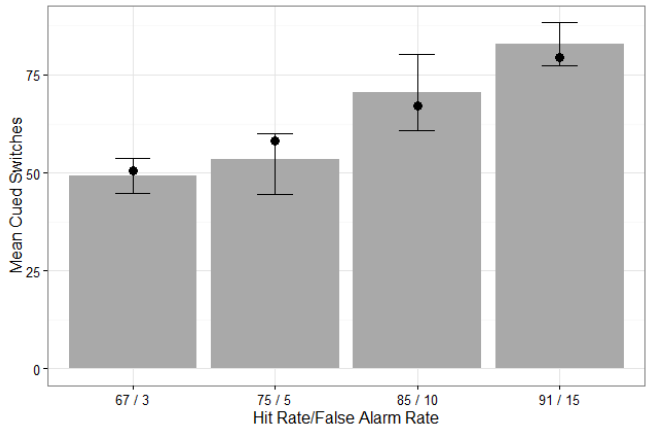


Figure 1 Bars depict empirical results. Error bars show a 95% confidence interval. Model fits are depicted by the black points.

We were also interested in analyzing the switching behavior when there was no automation cue. A one-way ANOVA revealed no significant differences in mean Uncued switches based on condition, F(3, 56) = .35, MSE = 1501.4, p > .05 (**Error! Reference source not found.**). This indicates that participants were not attuned to misses, if they were we would see an increasing trend of uncued switches as the hit rate decreased.

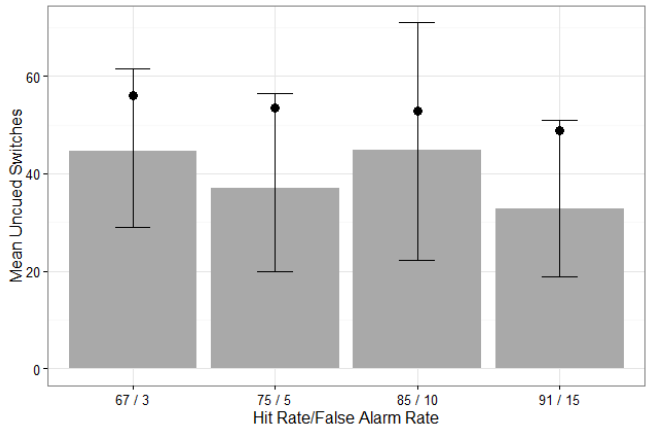


Figure 2 Bars depict empirical results. Error bars show a 95% confidence interval. Model fits are depicted by the black points.

We also explored how often participants responded to the alarm. The overall mean response rate to alarm was 0.968, and did not differ by condition (Table 3). The response rate results suggest that participants were merely responding when they heard the cue from the automation. Reaction time showed no effect by condition, F(3, 56) = .552, MSE = 119373, p > .05. This also suggests participants focused on the overall alarms, because if participants had focused on hits, they should have reacted faster overall when they heard the alarm, yet they did not. Finally, Ignored Cues showed no significant difference either, F(3, 56) = 1.02, MSE = 38.77, p > .05. . While null results cannot be interpreted strongly, this also suggests that participants were not impacted by FA, as we would expect to see an increasing trend in ignoring the cue as FA rates increased.

Table 3

*Mean Response rate and SD*

| Condition | Mean | SD |
|---|---|---|
| 91/15 | .99 | .02 |
| 85/10 | .94 | .15 |
| 75/5 | .95 | .2 |
| 67/3 | .99 | .02 |

Table 3 This table depicts the response rate to alarms in each condition.

## Process Model

*Description.* As mentioned earlier this model primarily took advantage of the utility learning mechanism in ACT-R. The model performs the same task as the participants. It also has to alternate between two windows that are only visible one at a time. It generally maintains attention on the primary screen, but it has two mechanisms for switching to the secondary screen. It can either decide to wait for the alarm and then switch, or it can decide to switch without hearing an alarm. Initiating a switch sets off a series of actions that lead to switching to the hidden secondary window. Once on the secondary screen it moves attention to the color box that represents the cart and responds accordingly. At this point if the cart is full, a reward is issued which affects the whole model.

The reward follows a differential propagation mechanism in which actions occurring more proximally in time receive a higher reward. This reward propagation is calculated using a mathematical formula which essentially works like the Rescorla-Wagner learning rule (Rescorla & Wagner, 1972). As such the decision to either switch or wait for the alarm receives a small reward when the secondary window displays a full cart, and it is in this way that the model learns to either to switch on its own or wait for the alarm. Details of the learning equation can be found in Fu & Anderson, (2006).

*Model Fit.* The model fit the Cued Switches data strongly, $R^2$ = .98 and RMSD = 3.5, and the Uncued Switches well, $R^2$ = .79 and RMSD = 13.4.

*Discussion.* The cognitive model helps us understand participants' behavior. The simple reward system changes the likelihood that different decision will be made as the model learns about the automation. In this case the cart being ready (red) rewarded whichever choice the model had made. In all the conditions the alarm was correct more often than not, as such the choice to wait for the alarm received more rewards and continued to be reinforced. Thus we see the same trend of cued switches in the model as in the participant data.

However, switching without a cue was also rewarded often enough that the model (and also the participants) continued to exhibit this behavior. The lack of cost is likely a part of the reason why we saw no tuning to the false alarm rate either. There was no significant penalty for the automation incorrectly cueing a switch, thus its mistakes did not affect participants trust enough to change their behavior. To look at the results another way, because there was no tangible cost to switching to the cart without a cue the model (and participants) continued to do so regardless of the automation characteristics.

## CONSIDERATIONS

The model also makes some interesting assumptions about the process used which warrant exploration. The current model does not use a declarative component in learning about the system. It is generally supported that as people engage in trust development they form memories and take previous experiences with the system into consideration when making judgments about trust (Lee & See, 2004). ACT-R is able to accumulate memories and based on frequency of use, it makes those memories more or less available (Anderson, 2007, pp. 95–104). However, the current model does not currently employ that module. The fact that we were able to get strong fits without explicitly modeling the memory component of trust does suggest that at least for this task it may not be part of the process. Another possibility is that memory for this task is a reflective component, i.e. participants only form explicit memories of their trust with regards to the automation after they are done with the task, in particular if they are asked about trust.

Another interesting issue concerns the difference between explicitly or implicitly communicating misses. In the current experimental design, participants are not explicitly notified of automation misses. Making miss information more explicit may result in more tuning to the miss behavior of the automation. If participants are attuned to misses, it would result in a pattern of increasing uncued switches as hit rate falls while maintaining a constantly high cued switch response.

We are currently exploring the effects of increasing the cost of switching to check the cart as we believe this to be the main driver for participants largely ignoring the automation behavior.

## ACKNOWLEDGEMENTS

This work was supported by a grant to JGT from the Office of Naval Research, The views and conclusions contained in this document do not represent the official policies of the US Navy.

## REFERENCES

Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press, USA.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the Independence of Compliance and Reliance: Are Automation False Alarms Worse Than Misses? *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(4), 564–572. doi:10.1518/001872007X215656

Fu, W.-T., & Anderson, J. R. (2006). From recurrent choice to skill learning: A reinforcement-learning model. *Journal of Experimental Psychology: General*, *135*(2), 184–206. doi:10.1037/0096-3445.135.2.184

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York: Wiley.

Gunzelmann, G., Byrne, M. D., Gluck, K. A., & Moore Jr, L. R. (2009). Using computational cognitive modeling to predict dual-task performance with sleep deprivation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(2), 251–260.

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. doi:10.1518/hfes.46.1.50_30392

Meyer, J. (2001). Effects of Warning Validity and Proximity on Responses to Warnings. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(4), 563–572. doi:10.1518/001872001775870395

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460.

Parasuraman, R., & Miller, C. A. (2004). Trust and Etiquette in High-Criticality Automated Systems. *Communications of the ACM*, *47*(4), 51–55.

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

Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical Conditioning II: Current Research and Theory* (pp. 64–99). Appleton-Century-Crofts.

Salvucci, D. D. (2006). Modeling Driver Behavior in a Cognitive Architecture. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *48*(2), 362–380. doi:10.1518/001872006777724417

Trafton, J. G., Altmann, E. M., & Ratwani, R. M. (2011). A memory for goals model of sequence errors. *Cognitive Systems Research*, *12*(2), 134–143. doi:10.1016/j.cogsys.2010.07.010

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: a synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. doi:10.1080/14639220500370105

Wiegmann, D. A., Rich, A., & Zhang, H. (2001). Automated diagnostic aids: The effects of aid reliability on users' trust and reliance. *Theoretical Issues in Ergonomics Science*, *2*(4), 352–367. doi:10.1080/14639220110110306

Yeh, M., & Wickens, C. D. (2001). Display Signaling in Augmented Reality: Effects of Cue Reliability and Image Realism on Attention Allocation and Trust Calibration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(3), 355–365. doi:10.1518/001872001775898269