COGNITIVE MODELING FOR HUMAN-AUTOMATION INTERACTION: A COMPUTATIONAL MODEL OF HUMAN TRUST AND SELF-CONFIDENCE

by

Katherine Jayne Williams

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

School of Mechanical Engineering
West Lafayette, Indiana
December 2021

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Neera Jain, Chair

School of Mechanical Engineering

Dr. Tahira Reid Smith

School of Mechanical Engineering

Dr. Brandon J. Pitts

School of Industrial Engineering

Approved by:

Dr. Nicole Key

To my family, friends, and mentors.

ACKNOWLEDGMENTS

Throughout my time at Purdue I have received a great deal of support. I would like to first express my gratitude to my advisor, Dr. Neera Jain. I am appreciative of her invaluable guidance and encouragement, and have grown personally and professionally through her leadership and expertise. I would also like to acknowledge the members of my advisory committee, Dr. Tahira Reid Smith and Dr. Brandon J. Pitts, for their support and valuable insights.

The work presented in this thesis was a part of a collaborative effort, and I would like to thank my colleagues. A special thanks to Madeleine Yuh, who has been my direct collaborator on this project and also a great friend. Additionally, I would like to thank the graduate students involved in the NSF CPS Frontier project for their insightful discussions: Adam, Maya, Sooyung, Vignesh, Albin, Jake, Dawei, Hansol, and Murat. The members of the Jain Research Lab have enriched my experience at Purdue. I am grateful to Jianqui (Jack), Trevor, Matt, Marcin, Michael, and Reynolds for their feedback, support, and friendship.

Finally, I would like to thank my parents and friends for their continued support throughout this journey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

x	Horizontal Position
y	Vertical Position
u	Participant Input
k	Trial Number
θ	Automation Assistance Level
Θ	Automation Assistance Level Set
Δt	Game Update Interval
ϕ	Game Wind Force
V	Speed
$s \in \mathcal{S}$	POMDP States
$a \in \mathcal{A}$	POMDP Actions
$o \in \mathcal{O}$	POMDP Observations
$b(s)$	Belief State
$\boldsymbol{\pi}$	Initial State Probability Function
$\mathcal{T}$	Transition Probability Function
$\mathcal{E}$	Emission Probability Function
	Subscripts/Superscripts
$\downarrow$	Denotes Low Value
$\uparrow$	Denotes High Value
$+$	Denotes Increase
$-$	Denotes Decrease
L	Low
M	Medium
H	High

ABBREVIATIONS

AUC	Area Under the Curve
FN	False Negative
FP	False Positive
FPR	False Positive Rate
HAI	Human-Automation Interaction
HMM	Hidden Markov Model
MDP	Markov Decision Process
NR	No Reliance
POMDP	Partially Observable Markov Decision Process
R	Reliance
ROC	Receiver Operating Characteristic
SC	Self-Confidence
srSC	Self-Reported Self-Confidence
T	Trust
TN	True Negative
TP	True Positive
TPR	True Positive Rate

ABSTRACT

Across a range of sectors, including transportation and healthcare, the use of automation to assist humans with increasingly complex tasks is also demanding that such systems are more interactive with human users. Given the role of cognitive factors in human decision-making during their interactions with automation, models enabling human cognitive state estimation and prediction could be used by autonomous systems to appropriately adapt their behavior. However, accomplishing this requires mathematical models of human cognitive state evolution that are suitable for algorithm design. In this thesis, a computational model of coupled human trust and self-confidence dynamics is proposed. The dynamics are modeled as a partially observable Markov decision process that leverages behavioral and self-report data as observations for estimation of the cognitive states. The use of an asymmetrical structure in the emission probability functions enables labeling and interpretation of the coupled cognitive states. The model is trained and validated using data collected from 340 participants. Analysis of the transition probabilities shows that the model captures nuanced effects, in terms of participants' decisions to rely on an autonomous system, that result as a function of the combination of their trust in the automation and self-confidence. Implications for the design of human-aware autonomous systems are discussed, particularly in the context of human trust and self-confidence calibration.

1. INTRODUCTION

The complexity of human interactions with autonomous systems is increasing, as evidenced in applications including intelligent transportation systems [1], autonomous vehicles [2], military operations [3], [4], and medical imaging systems [5]. In turn, this necessitates a greater understanding of these interactions and how they affect outcomes in terms of metrics such as performance [6]–[9]. It is well established that knowledge of a human's cognitive factors, or states, during their interactions with robots or other autonomous systems is vital to the design of effective human-automation interaction (HAI) [10], [11]. Indeed, cognitive factors such as a human's trust or self-confidence play a substantial role in their willingness, and decision, to rely on an autonomous system [12]–[18]. Therefore, designing autonomous systems that are responsive to the human's cognitive state could lead to improvements in task performance or human learning [18]. Specifically, models enabling cognitive state estimation and prediction could be used by autonomous systems to appropriately trigger system responses through methods such as transparency adaptation, automation behavior adaptation, and flexible autonomy [11]. However, accomplishing this requires mathematical models of human cognitive state evolution that are suitable for algorithm design.

Several conceptual frameworks have been proposed to model HAI and specifically the role of various cognitive factors in human behavior and decision-making [14], [16], [17], [19]–[22]. A majority of these frameworks are centered around human *trust* in automation [16], [17], [21], [22] which is well established as a fundamental psychological factor that can be defined in an HAI context as the belief that the automation will help the human achieve their goals in an uncertain situation [14]. Moreover, in HAI scenarios that involve a learning context, humans are also affected by their self-confidence [13] which enhances motivation to improve task performance when learning [23]. Importantly, researchers agree that human reliance on automation is coupled to both human trust and self-confidence [13], [24]–[28].

An overview of computational models of human trust or self-confidence are summarized in Table 1.1 and shows that a limited number of models exist that incorporate both states [13], [15], [40]. Many of these models are based upon the *'confidence vs trust' hypothesis*, originally developed in [13], that assumes a human's reliance on a given system is dependent

Table 1.1. Summary of computational models of trust and self-confidence. *denotes models that use the *'confidence vs trust' hypothesis*

	Category				
Papers	Trust	Self-Confidence	T-SC Coupling	Probabilistic	Deterministic
Lee and Moray, 1992 [24]	✓				✓
Lee and Moray, 1994* [13]	✓	✓			✓
Gao and Lee, 2006* [15]	✓	✓			✓
Maanen et al., 2011 [29]	✓				✓
Mikulski et al., 2012 [30]	✓			✓	
Saeidi et al., 2015* [31]	✓	✓			✓
Juvina et al., 2015 [32]	✓			✓	
Xu and Dudek, 2015 [33]	✓			✓	
Floyd et al., 2015 [34]	✓				✓
DeVisser et al., 2018 [6]	✓				✓
Chen et al., 2018 [35]	✓			✓	
Sadrfaridpour et al., 2018* [36]	✓	✓			✓
Wagner et al., 2018 [37]	✓				✓
Tao et al., 2020 [38]		✓		✓	
Azevedo-Sa and Yang, 2021 [39]	✓				✓
This thesis	✓	✓	✓	✓	

on a difference between the human's trust in the autonomous system and confidence in their ability to execute the task manually. For example, this hypothesis states that a person whose self-confidence exceeds their trust in the automation will choose to perform the task manually, and vice versa. However, some researchers have published results that contradict this hypothesis [41], [42]. For example, in [42], the authors show that in a signal detection task, despite their trust in the system being lower than their self-confidence, participants still relied on the system instead of completing the task manually. Furthermore, the authors of [41] suggest that operators who have both high trust and high self-confidence tend to prefer a higher level of automation. Therefore, further investigation of the coupling between trust and self-confidence is needed to characterize how different combinations of these cognitive states affect human reliance decisions and subsequent performance. To the knowledge of the author, existing models do not mathematically characterize this coupling.

The primary contribution of this thesis is a probabilistic discrete-state model of human trust and self-confidence dynamics as they relate to a human's repeated interactions with an autonomous system. An important feature of the model is its interpretability, which is achieved by first defining a model structure grounded in cognitive psychology and human factors literature, and then parameterizing it using human subject data collected in the context of a game-based task. The model considers coupling between the states themselves, as well as coupling between the human's reliance on the autonomous assistance and the cognitive states. Furthermore, the model leverages both behavioral and self-report data for model parameter estimation, collected from 340 human subjects. It is shown that the model's predictions are consistent with the findings of [41], [42] in that the 'confidence vs. trust' hypothesis does not account for all scenarios of trust and self-confidence interactions. Instead, the coupled effect of human trust and self-confidence on reliance is captured by the state transition probabilities of the trained model and underscores the need for computational models that can be used for algorithm design for improved HAI.

The thesis is organized as follows. In Chapter 2, the formulation of the trust and self-confidence modeling framework is presented. The human subject study, including experimental design and implementation, is outlined in Chapter 3. The modeling, training, and validation process is discussed in Chapter 4. The trained model is analyzed in Chapter 5, followed by a discussion of the implications of the results on the design of human-responsive automation. Finally, conclusions and future research directions are discussed in Chapter 6.

2. MODEL DEFINITION

A partially observable Markov decision process (POMDP) is an extension of a Markov decision process (MDP) and is defined as a 7-tuple, $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{E}, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, and $\mathcal{O}$ is a finite set of observations [43]. The transition probability function $\mathcal{T}$ governs the transition from the current state s to the next state s', given the action a. The emission probability function $\mathcal{E}$ governs the likelihood of observing o, given that the process is in state s. Finally, the reward function $\mathcal{R}$ and discount factor γ can be used to synthesize an optimal control policy but will not be used in the scope of this work. A POMDP accounts for observability through hidden states; this is particularly useful in the modeling of human cognitive dynamics which cannot always be directly measured or observed. This structure is leveraged here to establish a gray-box modeling framework for estimation and prediction of human trust and self-confidence that can be parameterized using human subject data. This promotes interpretability of the model. The model definition is supported by existing literature establishing key relationships between the cognitive states of interest, available observations, and relevant actions as discussed next.

First, the set of states $\mathcal{S}$ is defined as tuples containing the *Trust* state s_T and the *Self-Confidence* state s_{SC}, in which each state is attributed either a low ($\downarrow$) or high ($\uparrow$) value. This discrete state definition has been employed in prior POMDP models of human cognitive states and was shown to be sufficient for real-time trust calibration [44]. Given the literature discussed in Chapter 1 citing the coupling between human trust and self-confidence, the states are assumed to be coupled according to the following transition probability functions: $\mathcal{T}(s'_T|s_T, s_{SC})$ and $\mathcal{T}(s'_{SC}|s_T, s_{SC})$. Next, the set of actions $\mathcal{A}$ is defined as those variables that affect the state evolution. For HAI contexts, this includes the automation input (to their interaction with the human) as well as the human's experience with the automation. The latter is characterized here as the system performance which reflects the combined performance of the human-automation team at the previous time index. Therefore, $\mathcal{A}$ consists of tuples containing the *Automation Input* a_A and the *Performance* a_P.

Finally, the set of observations $\mathcal{O}$ is defined as the observable characteristics of the human's decision. As discussed earlier, it is well established in literature that human reliance on automation is affected by both the human's trust in the automation as well as their self-confidence [19]. Therefore, the emission probability function for reliance is defined as $\mathcal{E}(o_R|s_T, s_{SC})$. However, while a POMDP can be trained with fewer observations than states, doing so makes interpretation of the states difficult. Instead, self-reported self-confidence is used as a second observation for estimating the human's self-confidence state; this is described by the following emission probability function: $\mathcal{E}(o_{srSC}|s_{SC})$. The use of self-reported self-confidence here is supported by its use in work concerning the application of intelligent tutoring system (ITS) automation to train a self-confidence model [38]. This creates asymmetry in the emission probability function that aids interpretability of the model, as discussed in Chapter 4.

The proposed POMDP model definition is summarized in Table 2.1 and depicted in Fig. 2.1.

Table 2.1. Definition of the human trust—self-confidence (T-SC) POMDP model. Human trust and self-confidence are modeled as hidden states. The hidden states are affected by actions corresponding to the user's performance and the input provided by the automation. The observable characteristics of the user's chosen reliance and self-reported self-confidence are modeled as the observations of the POMDP.

States s $\in \mathcal{S}$	$\mathcal{S} = \begin{bmatrix} \text{Trust } s_T \\ \text{Self-Confidence } s_{SC} \end{bmatrix}$	$s_T \in T$ $T = \left\{ \begin{matrix} \text{Low Trust } T\downarrow \\ \text{High Trust } T\uparrow \end{matrix} \right\}$
		$s_{SC} \in SC$ $SC = \left\{ \begin{matrix} \text{Low Self-Confidence } SC\downarrow \\ \text{High Self-Confidence } SC\uparrow \end{matrix} \right\}$
Actions a $\in \mathcal{A}$	$\mathcal{A} = \begin{bmatrix} \text{Performance } a_P \\ \text{Automation Input } a_A \end{bmatrix}$	$a_P \in P$ $P = \left\{ \begin{matrix} \text{Performance Deterioration } P^- \\ \text{Performance Improvement } P^+ \end{matrix} \right\}$
		$a_A \in A$ $A = \{\text{Context Specific}\}$
Observations o $\in \mathcal{O}$	$\mathcal{O} = \begin{bmatrix} \text{Reliance } o_R \\ \text{Self-Reported Self-Confidence } o_{srSC} \end{bmatrix}$	$o_R \in R$ $R = \left\{ \begin{matrix} \text{No Reliance } R_{NR} \\ \text{Reliance } R_R \end{matrix} \right\}$
		$o_{srSC} \in srSC$ $srSC = \left\{ \begin{matrix} \text{Low Self-Confidence } srSC\downarrow \\ \text{High Self-Confidence } srSC\uparrow \end{matrix} \right\}$

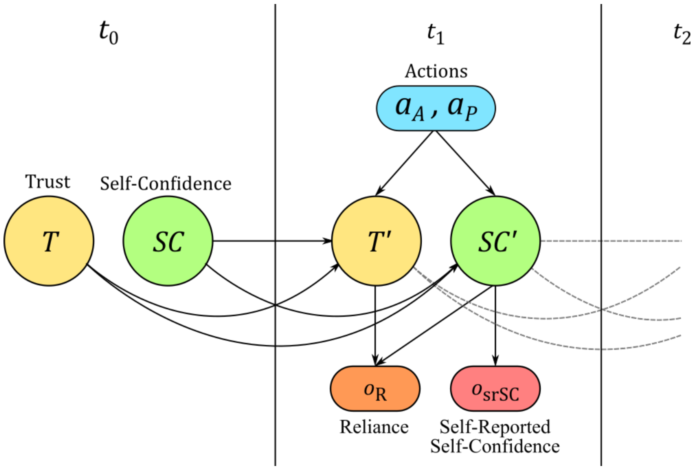


Figure 2.1. A representation of the proposed POMDP model of trust and self-confidence. The transition probabilities of trust and self-confidence are dependent on both of the previous states of trust and self-confidence. The reliance observation is dependent on both the trust state and self-confidence state. The self-reported self-confidence observation is dependent on only the self-confidence state.

Using the transition and emission probabilities, the probability distribution over the states, otherwise known as the belief state $b(s)$, can be calculated using Equation 2.1, in which $P(\cdot)$ denotes probability.

$$b'(s') = P(s'|o, a, b(s)) = \frac{P(o|s', a) \sum_{s \in S} P(s'|s, a)b(s)}{\sum_{s' \in S} P(o|s', a) \sum_{s \in S} P(s'|s, a)b(s)} \tag{2.1}$$

3. HUMAN SUBJECT STUDY

In Section 3.1, the design and intent of the human subject study for model training data collection is described. The implementation of the study is discussed in Section 3.2 and analysis of behavioral and self-report data collected from the experiment is presented in Section 3.3.

3.1 Experiment Design

Human subject data is collected in the context of a simple game-based task to parameterize the human trust—self-confidence (T-SC) model. The experimental platform is an online obstacle avoidance game in which participants must perform the task of maneuvering an avatar (depicted as a penguin) across the screen in the shortest amount of time while avoiding collisions with obstacles. However, participants are informed that an autonomous assistant is also available to help them play the game. The autonomous assistant scales the user's mouse input by a parameter θ. The extent of the scaling is subdivided into three sets of factors: $\Theta_L \in \{0.7, 0.8, 0.9\}$, $\Theta_M \in \{1.0, 1.1, 1.2\}$ and $\Theta_H \in \{1.3, 1.4, 1.5\}$, where $\theta \in \Theta_{\text{j}}$ for $\text{j} \in \{L, M, H\}$. In particular, when $\theta < 1$ the user will experience an attenuation of their mouse input, and when $\theta > 1$ their input will be amplified. In order to obtain training data that is agnostic to the dynamics of a specific autonomous assistance algorithm, the value of θ experienced by each participant is assigned to them according to the between-subjects experiment design described below. The goal of the experiment is to obtain a set of training data that captures the effect of a range of values of the autonomous assistant's input on participants' behavior. Whether a particular value of θ helps or hinders the participant is a function of their skill level.

In the game, the penguin avatar moves at a constant speed, and its position is controlled by the participant's mouse movement. The penguin's x and y position are governed by the following dynamical equations:

$$\begin{aligned} x_{t+1} &= x_t + \Delta t V cos(\theta_k u_t) + \phi \\ y_{t+1} &= y_t + \Delta t V sin(\theta_k u_t) \end{aligned} \tag{3.1}$$

where $[x_t, y_t]^T \in \mathbb{R}^2$ are the penguin's position at time t, $u_t \in \mathbb{R}$ is the participant's (mouse) input, and $\theta_k \in \mathbb{R}$ is the scaling factor provided by the autonomous assistant in the k^{th} trial for $k = 1, \ldots, 10$. The game update discrete time interval is Δt, V is the constant speed, and ϕ is an added "wind" effect which increases in the upward vertical direction. Table 3.1 provides the specific parameter values used in the experiment.

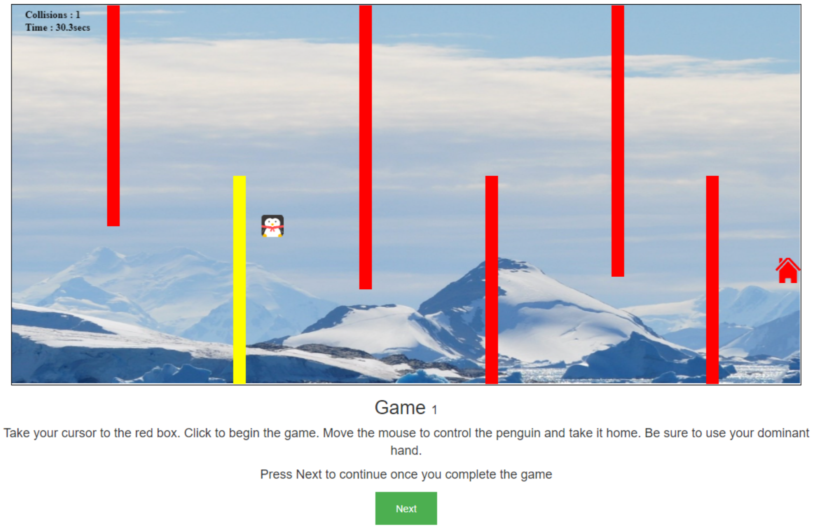
Table 3.1. Game parameters

Parameter	x_0	V	Δt	θ_0	ϕ
Value	[0, 200]	75 pixel/sec	0.02 sec	1	bottom $\{0.75\}$ middle $\{1.25\}$ top $\{1.75\}$

A between-subjects study is designed to elicit changes in each participant's trust in the autonomous assistant *and* confidence in their ability to play the game (i.e their self-confidence) over the course of 10 game trials. Fig. 3.3 shows the sequence of events for each trial in the experiment. Participants are asked to decide whether to rely or not rely on the autonomous assistant prior to every trial, as shown in Fig. 3.2a. Regardless of their reliance choice, prior to the first trial, each participant is randomly assigned to one of the three Θ sets. Then, for their first 5 trials, a single θ_1 value is randomly selected within the given Θ set. In this way, each participant experiences a constant input from the autonomous assistant for 5 repeated trials. Note that the participant is not informed of the specific θ value that is being applied to their input; they only know that the autonomous assistance is available and that they can turn it on or off. Moreover, for any game trial that they choose not to rely on the autonomous assistant, $\theta_k = 1$. After each trial, participants are provided a definition of trust and self-confidence before being prompted to rate their trust (in the autonomous assistant) and self-confidence on a numerical scale of 0-100. This is shown in Fig. 3.2b.

At the 6th trial, a step change in the Θ set is introduced. The purpose of this step change is to further perturb the participant's trust and self-confidence. Note that to avoid too large of a step change for some participants relative to others, no participant for whom $\theta_1 \in \Theta_L \vee \Theta_H$ experiences $\theta_2 \in \Theta_L \vee \Theta_H$. The choice of introducing the step change after 5 trials was based on data collected through pilot experiments. For the remaining five trials, a single θ_2 value is then randomly selected within the new Θ set. Equation 3.2 describes the step change in the θ value given the progression of k trials, with the caveat that $\theta_k = 1$ for any trial k during which the participant chooses not to rely on the autonomous assistant.

$$\theta_k = \begin{cases} \theta_1, & k \leq 5 \\ \theta_2, & 5 < k \leq 10 \end{cases} \tag{3.2}$$



Collisions : 1
Time : 30.3secs

Game 1

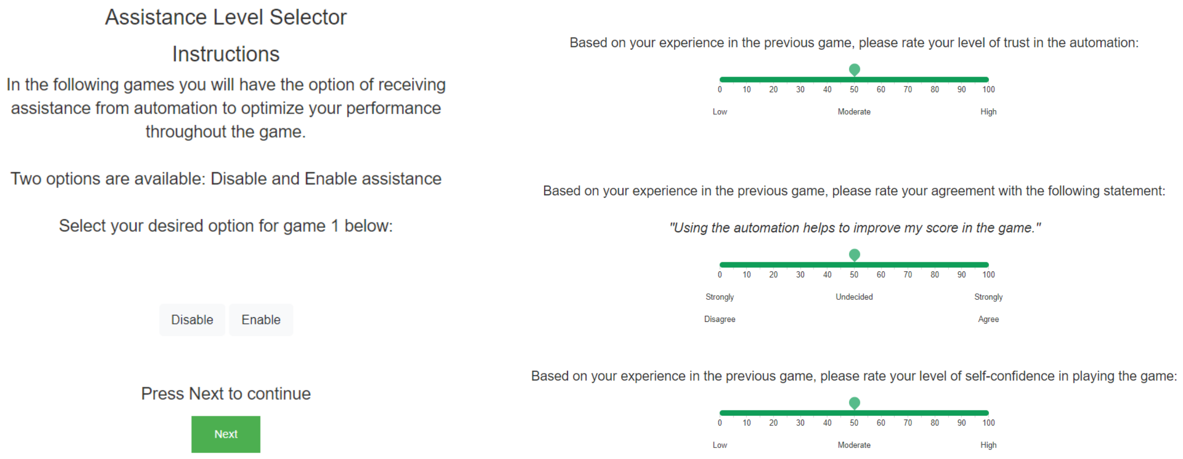
Take your cursor to the red box. Click to begin the game. Move the mouse to control the penguin and take it home. Be sure to use your dominant hand.

Press Next to continue once you complete the game

Figure 3.1. A screenshot of the web-deployed experiment platform in which the participant must guide a penguin across the game screen to its home while avoiding obstacles placed in its path.

3.2 Implementation

A total of 367 individuals participated in, and completed, the study. These participants were recruited from the Amazon Mechanical Turk platform [45] and completed the study online. To ensure the collection of quality data, the following criteria were applied to participant selection: participants must reside in the United States, have completed more than 500 Human Intelligence Tasks (HITs), and have a minimum HIT approval rate of 95%. Each participant provided their consent electronically and was compensated US$1.34 for their participation. The Institutional Review Board at Purdue University approved the study. Due to



Assistance Level Selector

Instructions

In the following games you will have the option of receiving assistance from automation to optimize your performance throughout the game.

Two options are available: Disable and Enable assistance

Select your desired option for game 1 below:

Disable Enable

Press Next to continue

Next

Based on your experience in the previous game, please rate your level of trust in the automation:

0 10 20 30 40 50 60 70 80 90 100

Low Moderate High

Based on your experience in the previous game, please rate your agreement with the following statement:

"Using the automation helps to improve my score in the game."

0 10 20 30 40 50 60 70 80 90 100

Strongly Disagree Undecided Strongly Agree

Based on your experience in the previous game, please rate your level of self-confidence in playing the game:

0 10 20 30 40 50 60 70 80 90 100

Low Moderate High

(a) Reliance Selection Page

(b) Survey Page

Figure 3.2. Example screenshots of the questions participants answer after each trial of the web-deployed experiment platform. (a) The reliance selection page in which participants are asked to select to either disable or enable the automation assistance. (b) The survey questions in which participants are asked to rate their trust and self-confidence on a numerical scale from 0-100.

the online nature of the experiment, and given lack of participant supervision, it is assumed that some participants were not adequately engaged in the study. This was reflected in their unusually low game completion time and high rate of collisions. To remove any outlying participants, the data from participants with at least three trials in which their game times were below the 25 percentile and with four or more collisions were filtered. These conditions were chosen because they indicated that the participant dragged the penguin across the screen without attempting to avoid the obstacles. As a result, 27 participants were removed from the data set. The resulting data set consists of 340 participants from the United States (145 females, 190 males, five preferred not to disclose or did not identify within either gender), ranging in age from 18-77 (mean 39.0 and standard deviation 11.9, two participants did not disclose age).

3.3 Behavioral and Self-Reported Data

Prior to training the model, the self-reported data is analyzed to identify behavioral trends. First, each participant's trust and self-confidence is identified as high or low by

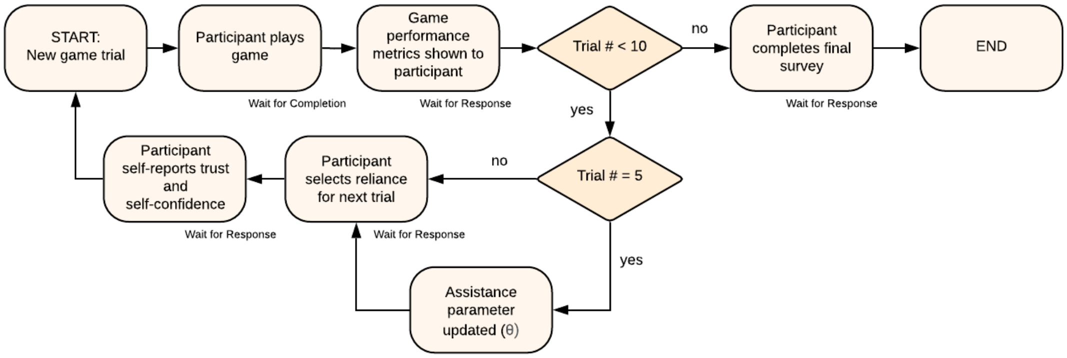

START:
New game trial
Participant plays game
Wait for Completion
Game performance metrics shown to participant
Wait for Response
Trial # < 10
no
Participant completes final survey
Wait for Response
END
yes
Trial # = 5
no
Participant selects reliance for next trial
Wait for Response
Participant self-reports trust and self-confidence
Wait for Response
yes
Assistance parameter updated (θ)

Figure 3.3. The sequence of events in the experiment. The participant completes a practice trial prior to completing ten trials of the game.

comparing the participant's self-reported value to the 50th percentile from all data. In Fig. 3.4 the mean value of the number of collisions across all data points pertaining to each self-reported state combination is used to plot the average collisions. The number of instances in which participants chose to rely is counted and divided by the total number of data-points in each self-reported state combination to find and plot the reliance rates. There exist clear distinctions between each cognitive state and the number of collisions and chosen reliance level of each participant associated with their reporting of each state. From Fig. 3.4, it can be seen that state combinations, such as $T{\downarrow}SC{\downarrow}$ and $T{\uparrow}SC{\downarrow}$, correspond to poorer performance. The established relationship between trust and reliance captured in previously published trust models is further underscored in Fig. 3.4. When trust is high, the reliance rate is high, and vice versa. However, the addition of self-confidence affects the user's likelihood to rely on the autonomous assistant. When trust is low, the users with low self-confidence are 12% more likely to rely on the autonomous assistant than those with high self-confidence. It should also be noted that when $T{\uparrow}SC{\uparrow}$, it would have been expected that users would not rely on the assistant as often. Instead, participants who reported being in the $T{\uparrow}SC{\uparrow}$ state demonstrated a high reliance rate and low number of collisions. Finally, the data show an almost inverse relationship between the $T{\uparrow}SC{\uparrow}$ and $T{\downarrow}SC{\downarrow}$ states. These findings will be used to aid in model state sorting, as discussed in the next section.

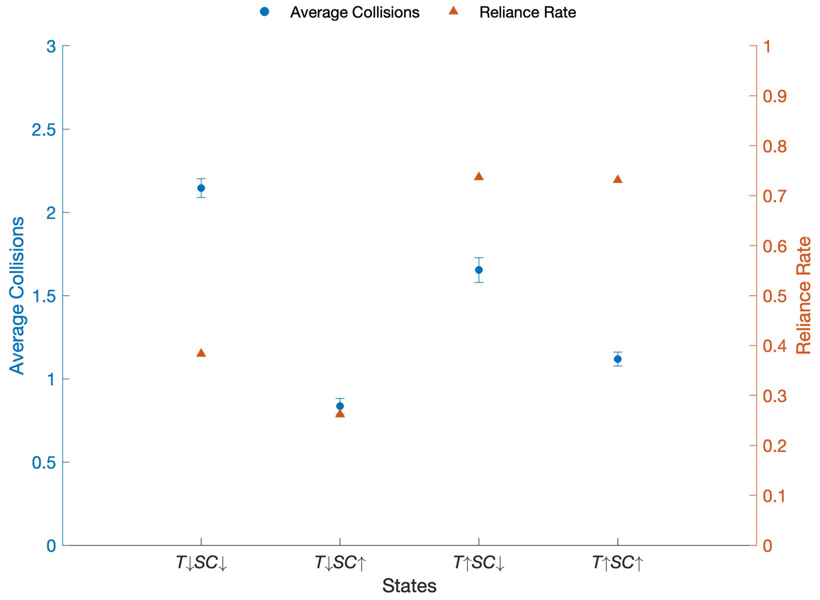


Figure 3.4. Average collisions (left y-axis) and reliance rate (right y-axis) corresponding to the four combinations of trust and self-confidence, $T{\downarrow}SC{\downarrow}$, $T{\downarrow}SC{\uparrow}$, $T{\uparrow}SC{\downarrow}$, and $T{\uparrow}SC{\uparrow}$, as self-reported by participants. The error bars of the average collisions represent the standard error of the mean across participants.

4. MODEL TRAINING AND VALIDATION

The adaptation of the model to the specific HAI context considered in this thesis is first discussed in Section 4.1. This is followed by a description of the methods used for model training (Section 4.2) and model validation (Section 4.3).

4.1 Model Definition

Recall the T-SC cognitive state model defined in Table 2.1. In the context of the experimental platform used for data collection, there are two relevant performance metrics: the number of collisions between the penguin and the obstacles, and the time taken to navigate the penguin to its home in the game environment. Therefore, the performance action is further divided into tuples containing the number of *Collisions* a_C and *Game Time* a_G, as shown in Equation 4.1. Additionally, the automation input a_A is the assistance value θ, discretized into the sets Θ_L, Θ_M and Θ_H as described in Chapter 3 and referenced in Equation 4.2.

$$\begin{aligned} a_C \in C = \{\text{Collision Decrease } C^-,\ \text{Collision No Change } C^0,\ \text{Collision Increase } C^+\} \\ a_G \in G = \{\text{Game Time Decrease } G^-,\ \text{Game Time Increase } G^+\} \end{aligned} \tag{4.1}$$

$$a_A \in A = \{\Theta_L,\ \Theta_M,\ \Theta_H\} \tag{4.2}$$

The transition probabilities for trust $\mathcal{T}_T : \mathcal{S} \times T \times \mathcal{A} \rightarrow [0,1]$ and self-confidence $\mathcal{T}_{SC} : \mathcal{S} \times SC \times \mathcal{A} \rightarrow [0,1]$ are each represented by $4 \times 2 \times 18$ matrices that map the probability of transitioning from combinations of states $\mathcal{S}$ of trust $s_T \in T$ and self-confidence $s_{SC} \in SC$ to the next states of trust and self-confidence, respectively, given an action $a \in \mathcal{A}$. The state combination transition probabilities are the product of the individual transition probabilities of trust and self-confidence, as given by

$$\mathcal{T}(s'|s,a) = \mathcal{T}(s'_T|s_T, s_{SC}, a)\mathcal{T}(s'_{SC}|s_T, s_{SC}, a)\ \ . \tag{4.3}$$

The emission probability function for reliance $\mathcal{E}_R : \mathcal{S} \times R \rightarrow [0,1]$ is represented by a 4×2 matrix that maps the probability of reliance on automation $o_R \in R$ given the current trust and self-confidence belief states. The emission probability function for self-reported self-confidence $\mathcal{E}_{srSC} : SC \times srSC \rightarrow [0,1]$ is represented by a 2×2 matrix that maps the probability of low or high self-reported self-confidence $o_{srSC} \in srSC$ given the current self-confidence state. The overall emission probabilities are the product of the individual reliance and self-reported self-confidence emission probabilities, given by

$$\mathcal{E}(o|s) = \mathcal{E}(o_R|s_T, s_{SC})\mathcal{E}(o_{srSC}|s_{SC}) \ . \tag{4.4}$$

Finally, the initial state probabilities for trust $\boldsymbol{\pi}_T : 1 \times T \rightarrow [0,1]$ and self-confidence $\boldsymbol{\pi}_{SC} : 1 \times SC \rightarrow [0,1]$ are both represented by 1×2 matrices that represent the probability of the initial trust state s_T and self-confidence state s_{SC} respectively. As shown in Fig. 2.1, the reliance observation is dependent on both the current trust and self-confidence states. However, the self-reported self-confidence observation is only dependent on the current self-confidence state. In total, there are 153 effective parameters from the 18 combinations of actions, 4 combinations of states, and 4 observations.

4.2 Model Parameter Estimation

It is assumed that trust and self-confidence behavior for the general population can be represented by a common model. Therefore, the aggregated data of all participants is utilized in estimating the model parameters, resulting in 340 sequences of data. Previously, an extended version of the Baum-Welch algorithm was used to estimate the parameters of a discrete observation-space cognitive model [44]. However, literature suggests that the genetic algorithm is not as sensitive to the initialization of parameters and not as susceptible to local optima as compared to the Baum-Welch algorithm [46]. Therefore, the genetic algorithm in MATLAB's Optimization Toolbox [47] is implemented to optimize the parameters of the model to maximize the likelihood of the sequences given the model parameters. The forward algorithm is utilized to calculate the likelihood of the sequences [48] in which the algorithm computes, recursively over time, the joint probability of a state s_k given time k

and the series of observations $o_{1:k}$ and actions $a_{1:k}$ over time, i.e. $P(s_k, o_{1:k}, a_{1:k})$. The sum of $P(s_N, o_{1:N}, a_{1:N})$ is calculated to determine the likelihood of the sequence across all states at the end of the sequence at time N. This gives the probability of the action observation sequence, $P(o_{1:N}, a_{1:N})$.

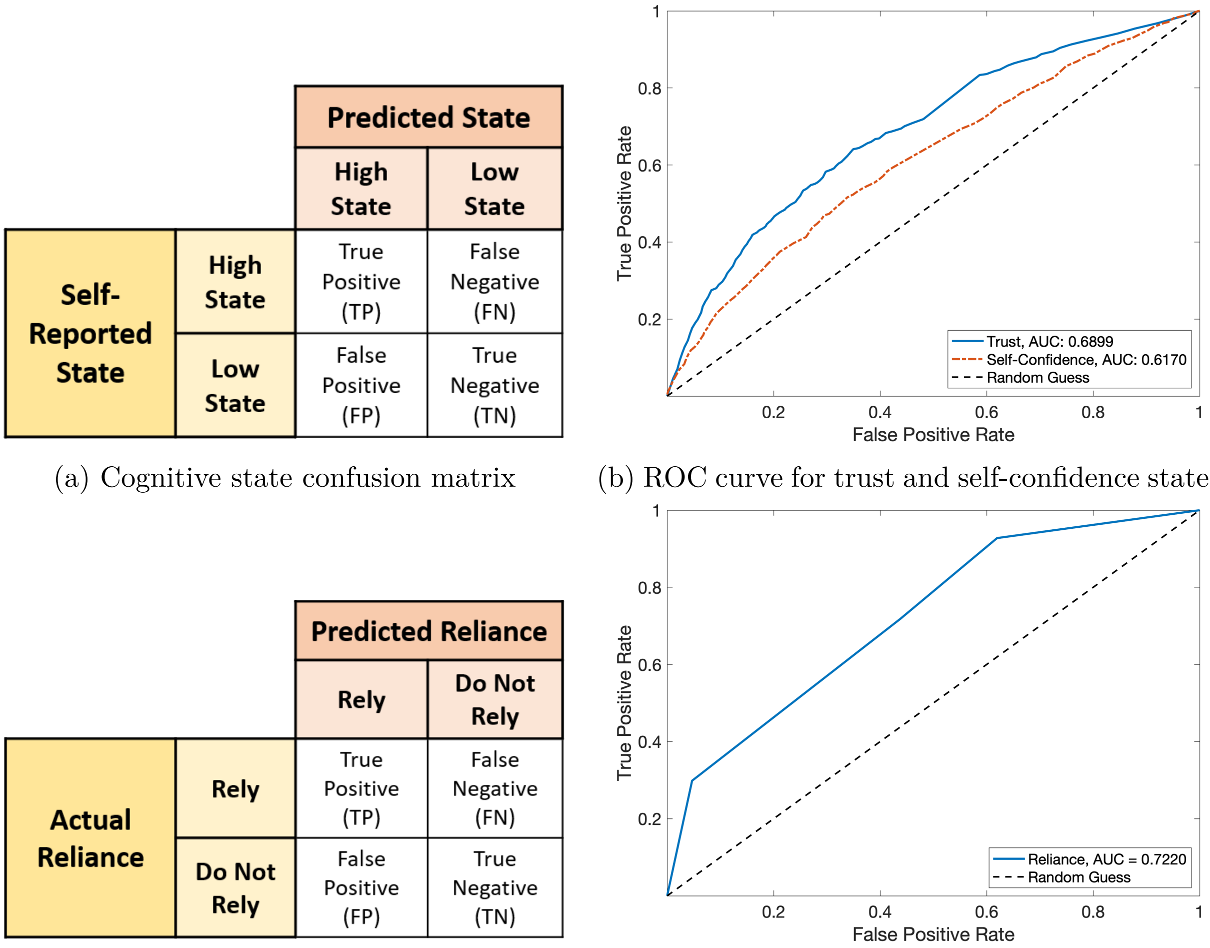
Prior to training the model, the order of the action combinations and observation combinations are established. However, this is not the case for the state combinations. The state combination order of the resulting transition, emission, and initial probability matrices is sorted into the order $T{\downarrow}SC{\downarrow}$, $T{\downarrow}SC{\uparrow}$, $T{\uparrow}SC{\downarrow}$, and $T{\uparrow}SC{\uparrow}$ after training the model by using established behavioral trends. Identifying the state combination of each row is possible due to the asymmetrical nature of the emission probability functions. The self-reported self-confidence emission probabilities are used to determine the self-confidence state order. The reliance emission probabilities are used to sort the trust state order by applying the well-known correlation between trust and reliance [24], [49]–[51]. After identifying the corresponding state combination of each row in the emission probability matrix, all rows and columns associated to states in the initial, transition, and emission probability matrices are re-ordered to match the prescribed state combination order.

4.3 Validation

To test the predictive capability of the model and check for over-fitting, two validations methods are used. A ten-fold cross validation is applied to the data in which the data is divided randomly into ten equal sets, or folds. The model is trained with 9 selected folds and validated using the 10^{th} fold. This is done ten times, in total, with each fold being used once for validation and the remaining 9 for training. The entire process is then repeated for ten iterations to increase the robustness of the validation log-likelihood values to variations in the training and testing data sets. One-way ANOVA tests between the ten iterations show that there is no statistical difference in the validation log-likelihoods ($\alpha = 0.05$, $p = 0.9341$). This indicates that the trained model has converged and is not over-fitting the data.

Next, receiver operating characteristic (ROC) curves are utilized to illustrate the performance of the model in predicting the cognitive states and chosen reliance of each participant.

The cognitive state ROC curves (Fig. 4.1b) are generated by comparing the self-reported cognitive states to the predicted belief state, as calculated by Equation 2.1, for all 340 participants' data. The belief state probability of high trust or self-confidence is first compared to a threshold probability, in which the predicted state is classified as high if the belief state probability is greater than the classification threshold probability. Then, the predicted state is compared to the self-reported state. As seen in Fig. 4.1a, this results in a true positive (TP), false positive (FP), true negative (TN), or false negative (FN), depending on if the predicted state is high or low and if the predicted state matches the self-report data. For classification thresholds of 0-100% in increments of 1%, this process is repeated for all data to find the true positive rate (TPR) and false positive rate (FPR) for each threshold probability. The TPRs and FPRs of each threshold are plotted, resulting in the ROC curve. The reliance ROC curve (Fig. 4.1d) is generated using a similar method, but instead, the maximum belief state probability is used to determine the corresponding emission probability. The emission probability is compared to a classification threshold probability to predict the participant's choice of reliance. TPRs and FPRs are found by comparing the predicted reliance to the participant's actual chosen reliance, as shown in Fig. 4.1c. The model can predict both cognitive state levels and reliance choice better than a random guess as shown in Figures 4.1b and 4.1d. This is further supported by the area under the curve (AUC), an aggregate performance measure across all thresholds. A higher AUC corresponds to a better model classification performance. The trained model achieves a trust AUC of 0.69, self-confidence AUC of 0.62, and reliance AUC of 0.72.



		Predicted State	
		High State	Low State
Self-Reported State	High State	True Positive (TP)	False Negative (FN)
	Low State	False Positive (FP)	True Negative (TN)

(a) Cognitive state confusion matrix

(b) ROC curve for trust and self-confidence states

		Predicted Reliance	
		Rely	Do Not Rely
Actual Reliance	Rely	True Positive (TP)	False Negative (FN)
	Do Not Rely	False Positive (FP)	True Negative (TN)

(c) Reliance confusion matrix

(d) ROC curve for reliance

Figure 4.1. Receiver Operating Characteristic (ROC) curves for cognitive state and reliance prediction. The given model classification performance is determined by the area under the curve (AUC), which is denoted in the legends of plots (b) and (d). As noted, the model achieves a trust AUC of 0.69, self-confidence AUC of 0.62, and reliance AUC of 0.72.

5. RESULTS AND DISCUSSION

In Section 5.1, the identified emission and transition probabilities are presented and interpreted in the context of the specific HAI scenario under consideration. This is followed by a discussion of the implications of the model for improving HAI (Section 5.2) and a review of limitations (Section 5.3).

5.1 Results and Analysis

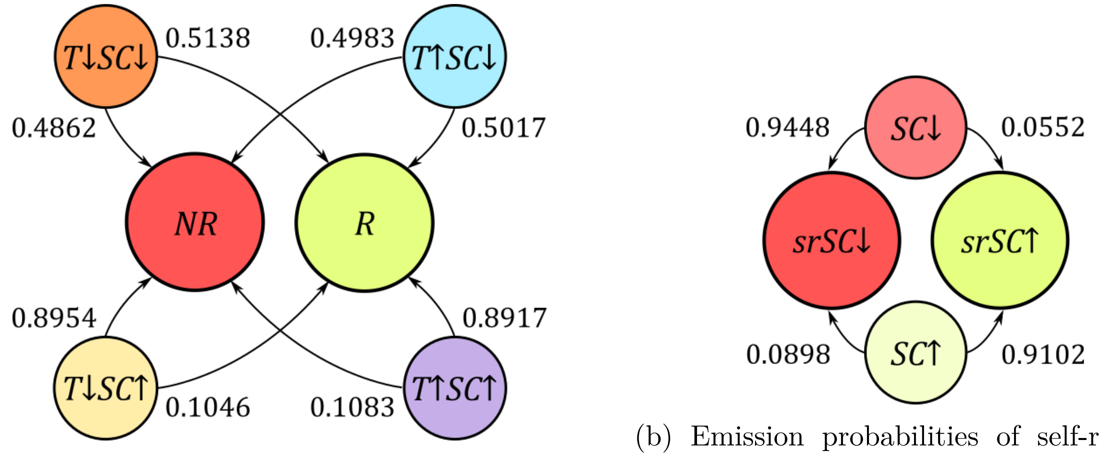
The results are presented and analyzed following the model structure, including the initial state probabilities (Section 5.1.1), emission probabilities (Section 5.1.2), and transition probabilities (Section 5.1.3).

5.1.1 Initial State Probabilities

A complete table of the initial state probabilities can be found in Appendix A.1. From Table A.1, it is inferred that participants tend to initially have high trust in the autonomous assistant (81.22%) and low self-confidence (60.70%). The initial high trust is consistent with existing literature that states that humans tend to have positivity bias towards automation, in which they trust automation prior to having any experience with it [52].

5.1.2 Emission Probabilities

Next the identified emission probabilities, visually depicted in Figs. 5.1a and 5.1b, are analyzed. Fig. 5.1b shows the probability of self-reported self-confidence given the the self-confidence state. The probabilities of the self-reported self-confidence being the same as the self-confidence states are 94.48% and 91.02% for low and high self-confidence respectively, suggesting that the state is capturing what the human perceives as their level of self-confidence. Next, Fig. 5.1a shows the probability of reliance given the trust and self-confidence states. The first observation to be made is that when self-confidence is high, the resulting probabilities behave similarly to the established trust and reliance relationship in which low and high trust lead to low and high reliance, respectively. For example, when



(a) Emission probabilities of reliance in which no reliance and reliance are denoted by NR and R, respectively.

(b) Emission probabilities of self-reported self-confidence in which high and low self-reported self-confidence are denoted by $srSC\uparrow$ and $srSC\downarrow$, respectively.

Figure 5.1. The emission probability function for reliance $\mathcal{E}(o_R|s_T, s_{SC})$ and self-reported self-confidence $\mathcal{E}(o_{srSC}|s_{SC})$. The probabilities are shown next to the arrows.

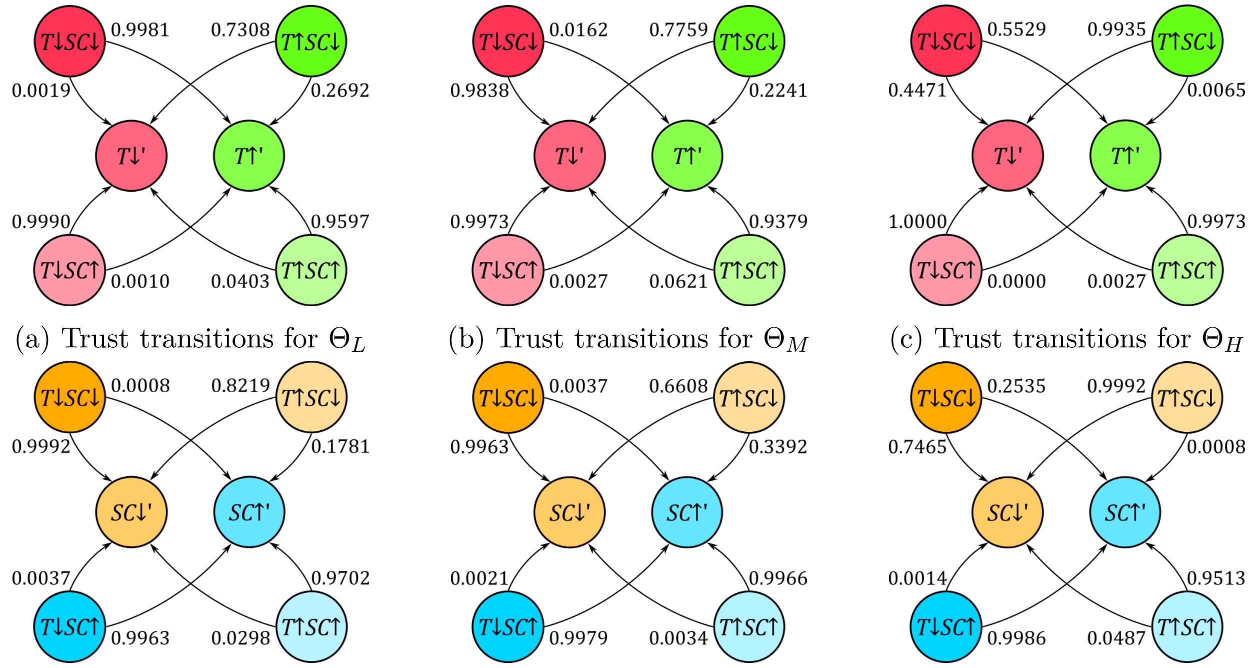
participants are in a state of low trust and high self confidence ($T{\downarrow}SC{\uparrow}$), they are highly likely (89.54%) to not rely on the automation, and when they are in the $T{\uparrow}SC{\uparrow}$ state, they are highly likely (89.17%) to rely. Interestingly, this relationship is not exhibited when self-confidence is low. Instead, when participants are in the $T{\downarrow}SC{\downarrow}$ state, the likelihood that they will disable (48.62%) or enable (51.35%) the automation assistance is nearly equally distributed. The same is true when they are in the $T{\uparrow}SC{\downarrow}$ state. This underscores the complex relationship between human trust and self-confidence in the context of HAI, which is further analyzed in the next subsection.

5.1.3 Transition Probabilities

Given that the POMDP consists of 3 discrete-valued actions that result in 18 distinct combinations of actions, there are a total of 18 different transition probability functions that describe the state transitions. The transition probability functions are divided to separate the probabilities of trust state transitions and probabilities of self-confidence state transitions. A complete review of all transition probabilities can be found in Appendix A.2. For clarity of exposition, a subset of these probabilities are analyzed here. Specifically, the actions associated with participants' performance—changes in the number of collisions and game time—are grouped into cases of performance improvement or deterioration, and the effect of the third action, the autonomous assistance, is analyzed within these groupings.

Overall Performance Improvement

The overall performance improvement case scenario is that in which the number of collisions decreases C^- and game time decreases G^-. When $a_A \in \Theta_L$, as shown in Figs. 5.2a and 5.2d, and for all state combinations, self-confidence is likely to remain the same at the next trial (>80%). Moreover, when the participant is in the $T{\downarrow}SC{\downarrow}$ state, they are very likely to transition to a state of high trust (99.81%), suggesting that *they associate performance improvement to the automation rather than themselves.* This is not the case for most participants in the $T{\uparrow}SC{\downarrow}$ state, though. Participants' cognitive state responses when they are in the $T{\uparrow}SC{\downarrow}$ state are similar for all a_A as shown in Figs. 5.2a-5.2f. They are likely to



(a) Trust transitions for Θ_L

(b) Trust transitions for Θ_M

(c) Trust transitions for Θ_H

(d) Self-confidence transitions for Θ_L

(e) Self-confidence transitions for Θ_M

(f) Self-confidence transitions for Θ_H

Figure 5.2. The transition probability function for trust $\mathcal{T}_T(s'_T|s_T, s_{SC}, a)$ and self-confidence $\mathcal{T}_{SC}(s'_{SC}|s_T, s_{SC}, a)$. The performance actions are the overall improvement case scenario in which the number of collisions decreases C^- and game time decreases G^-. The probabilities of transition are shown next to the appropriate arrows. (a) The trust transition probabilities for $a_A \in \Theta_L$. (b) The trust transition probabilities for $a_A \in \Theta_M$. (c) The trust transition probabilities for $a_A \in \Theta_H$. (d) The self-confidence transition probabilities for $a_A \in \Theta_L$. (e) The self-confidence transition probabilities for $a_A \in \Theta_M$. (f) The self-confidence transition probabilities for $a_A \in \Theta_H$.

transition to a state of low trust (73.08%, 77.59%, 99.35%) while they are likely to remain in a state of low self-confidence (82.19%, 66.08%, 99.92%), suggesting that the decrease in trust may be a result of the user attributing the performance improvement more towards themselves than the automation. Upon closer analysis, when $a_A \in \Theta_L \vee \Theta_M$, participants had a 26.92% and 22.41% chance, respectively, of remaining in a state of high trust, and a 17.81% and 33.92% chance, respectively, of transitioning to a state of high self-confidence. Therefore the different values of a_A may result in different attributions of performance between the user and automation which then affect the participants' cognitive state responses. When $a_A \in \Theta_H$, as shown in Figs. 5.2c and 5.2f, and when the participant is in the $T{\downarrow}SC{\downarrow}$ state, the probability of them transitioning to a state of high trust (55.29%) or remaining in a state of low trust (44.71%) are approximately equally distributed. On the other hand, they are more likely to remain in a state of low self-confidence (75%) than to transition to a state of high self-confidence. These participants may associate the cause of performance improvement slightly more with the automation than themselves.

Interestingly, for all a_A, when participants are in a state of high self-confidence and experience an overall improvement in performance, they are very likely to remain in a state of high self-confidence as well as maintain the same level of trust in the autonomous assistant at the next trial. In other words *a participant's self-confidence affects their interpretation of their performance metrics, which in turn affects their trust in the automation.*

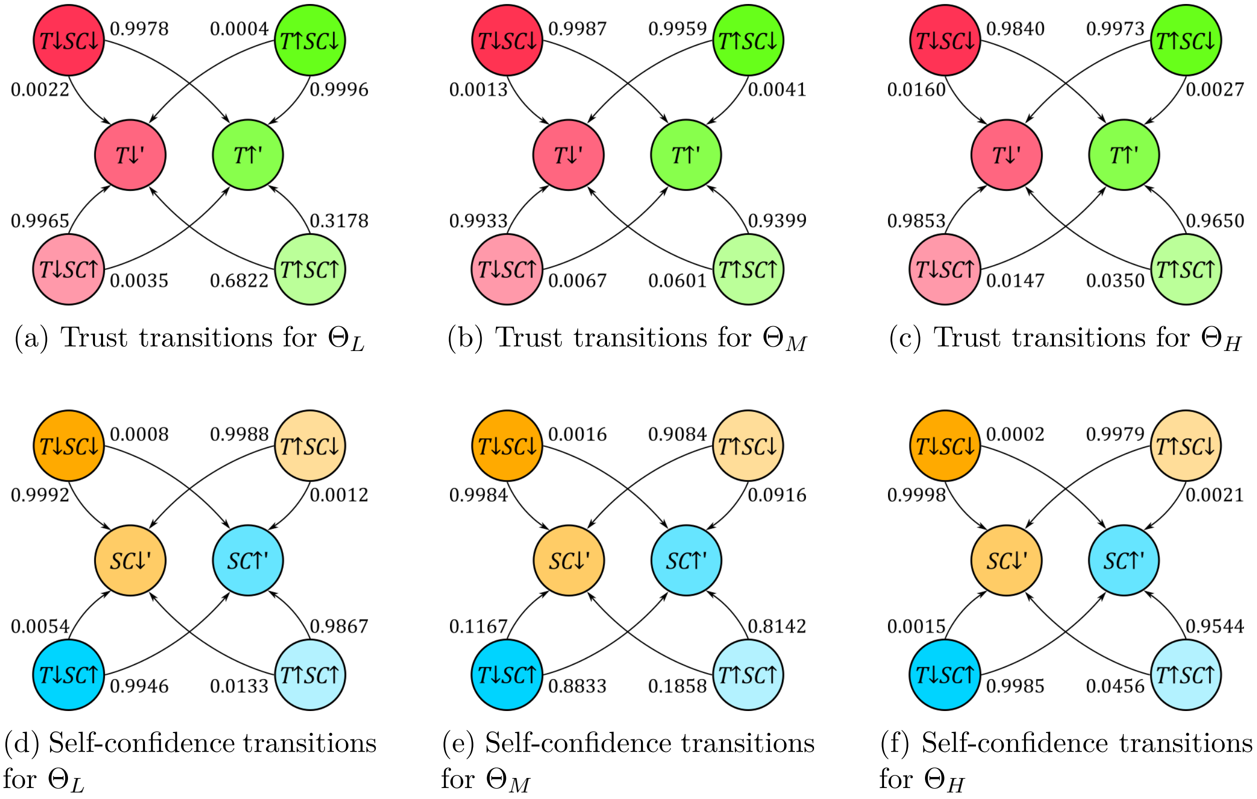
Partial Performance Improvement

For performance improvement, another case of interest is that in which the number of collisions does not change but the participants' game time decreases. This represents a case of partial improvement. When $a_A \in \Theta_L$, as shown in Table A.2 (see Appendix A.2), and when the participant is in the $T{\downarrow}SC{\downarrow}$ state, their likelihood of transitioning to a state of low trust (45.72%) or high trust (54.28%) is nearly equally distributed. However, they are likely to remain in a state of low self-confidence (79.49%). This is similar to when participants are in the $T{\uparrow}SC{\downarrow}$ state and $a_A \in \Theta_M$, as shown in Table A.3. When $a_A \in \Theta_H$, as shown in Table A.4, and the participant is in the $T{\downarrow}SC{\downarrow}$ state, they are highly likely (99.86%) to remain in a state of low self-confidence. However, their likelihood of transitioning to a state of high trust is only 29.52%. When $a_A \in \Theta_L \vee \Theta_H$ and participants are in the $T\!\downarrow\!SC\!\downarrow$ state, trust increasing suggests that they are attributing a slight improvement in performance to the automation rather than themselves. However, when $a_A \in \Theta_M$, the fact that participants in a state of high trust are equally likely to remain in their current state or transition to a state of low trust while their low self-confidence is likely to be maintained (84.12%) suggests that they are unsure of to whom they should attribute the improvement in performance.

In comparing these results to the overall improvement case, participants in a state of low self-confidence are still unlikely to gain confidence and transition to $SC\!\uparrow$, but they are now not as likely to attribute any improvement to the automation. This underscores the consequences, from the perspective of HAI, of a human being in a state of low self-confidence. *In other words, participants in a state of low self-confidence may have more difficulty in calibrating their trust in the automation than those with high self-confidence.* In turn this suggests that correct calibration of self-confidence is just as important as trust calibration, as discussed more in Section 5.2.

Overall Performance Deterioration

Next, cases in which participants' performance deteriorates between game trials is analyzed. For all a_A, when performance *deteriorates* and participants are in the $T\!\downarrow\!SC\!\downarrow$ state, their trust is highly likely to increase (99.78%, 99.87%, 98.40%) at the next trial. However, they are likely to remain in a state of low self-confidence (99.92%, 99.84%, 99.98%). This suggests that these participants associate performance deterioration to themselves rather than the automation. On the other hand, the autonomous assistance input does have a greater effect on participants in states of high trust (either $T\!\uparrow\!SC\!\downarrow$ or $T\!\uparrow\!SC\!\uparrow$). When $a_A \in \Theta_M \vee \Theta_H$ (Figs. 5.3b and 5.3c), participants in a state of high trust are very likely (>90%) to transition to a state of low trust, regardless of their state of self-confidence. This suggests that they strongly attribute the decrease in performance to the autonomous assistant. This is not true when $a_A \in \Theta_L$, in which participants who are in a state of $T\!\uparrow\!SC\!\downarrow$ are likely to remain in a state of high trust at the next trial. These results highlight that while self-confidence affects



(a) Trust transitions for Θ_L

(b) Trust transitions for Θ_M

(c) Trust transitions for Θ_H

(d) Self-confidence transitions for Θ_L

(e) Self-confidence transitions for Θ_M

(f) Self-confidence transitions for Θ_H

Figure 5.3. The transition probability function for trust $\mathcal{T}_T(s'_T|s_T, s_{SC}, a)$ and self-confidence $\mathcal{T}_{SC}(s'_{SC}|s_T, s_{SC}, a)$. The performance actions are the overall deterioration case scenario in which the number of collisions increases C^+ and game time increases G^+. The probabilities of transition are shown next to the appropriate arrows. (a) The trust transition probabilities for $a_A \in \Theta_L$. (b) The trust transition probabilities for $a_A \in \Theta_M$. (c) The trust transition probabilities for $a_A \in \Theta_H$. (d) The self-confidence transition probabilities for $a_A \in \Theta_L$. (e) The self-confidence transition probabilities for $a_A \in \Theta_M$. (f) The self-confidence transition probabilities for $a_A \in \Theta_H$.

participants' attribution of changes in performance, so does the user's experience with the autonomous assistant.

Partial Performance Deterioration

Next, the case in which number of collisions does not change but the participants' game time increases is considered. For $a_A \in \Theta_L \vee \Theta_M \vee \Theta_H$, shown in Tables A.2, A.3, and A.4, respectively, and when participants are in the $T{\downarrow}SC{\downarrow}$ state, it is likely for their trust to increase (99.98%, 99.70%, 99.90%) at the next trial and likely for them to remain in a state of low self-confidence (95.12%, 99.76%, 100%). These results are consistent with those observed for the overall performance deterioration case. When $a_A \in \Theta_H$, however, and participants are in the $T{\uparrow}SC{\downarrow}$ state, their likelihood of transitioning to a state of low trust (57.68%) or high trust (42.32%) is more equally distributed than in the overall performance deterioration case. Therefore, the extent of the change in performance also affects participants' trust and self-confidence dynamics.

5.2 Implications on the Design of Human-Aware Autonomous Systems

As discussed in the previous section, depending on their performance and the input from the autonomous assistant, participants may attribute their successes and failures to either the automation or themselves. These observations are a demonstration of attribution theory, a theory concerned with the processes behind the attempts of humans to explain the cause of behaviors and events [53], [54]. Understanding the different attributions is important because reliance is not only affected by participants' beliefs about the automation's performance or reliability, but also by cognitive factors affecting this performance [14], in this case, participants' trust in the automation and self-confidence. Importantly, for the purpose of improving performance and safety outcomes for different HAI contexts, the proposed probabilistic model can be used to design cognitive state-based feedback policies that help human's correctly attribute changes in performance to themselves or the automation, and in turn better calibrate their trust in the automation and their self-confidence. Calibration of human trust in HAI is critical to preventing the pitfalls associated with humans under-trusting, or over-trusting autonomous systems. However, to date, less emphasis has been placed on calibration of self-confidence in HAI, despite the fact that a human who is incorrectly over-confident in their skills may under-trust the automation they are interacting with, and vice versa. The model analysis presented here shows that both states must be calibrated correctly for improving HAI. With knowledge of how the human's cognitive dynamics evolve, autonomous systems can be designed to facilitate this, for example, through the use of automation transparency.

5.3 Limitations

It is worthwhile to acknowledge some of the limitations of the proposed model for capturing human trust and self-confidence dynamics. It is assumed that the cognitive state dynamics evolve based on the *change* in the participant's performance rather than their absolute performance. In other words, in training the model, the behavior of a skilled participant who experienced slight improvement was not distinguished from that of a poor-performing participant who likewise had a slight performance improvement. In future work, this limitation can be mitigated by considering absolute performance in addition to the change in performance. Furthermore, as is the case with any model trained using human data, the conclusions drawn in this paper are specific to the HAI scenario under consideration. However, given the generalized definition of the POMDP states, observations, and actions, future work should investigate how well the transition and emission probability functions translate to other HAI scenarios and the extent to which new human data is needed for doing so.

Finally, while a POMDP modeling framework was chosen here for several benefits it offers in capturing the probabilistic nature of human cognitive dynamics, a limitation of POMDPs is their scalability. Modest increases in the numbers of actions, states, or observations can lead to parameter explosion, thereby increasing the amount of data needed for parameter estimation. Therefore, the proposed framework may not scale well to more complex HAI scenarios in which additional actions may need to be defined, for example, to capture the nature of the automation's input. Similarly, further discretizing the trust or self-confidence states beyond two discrete values will also lead to increased model complexity. Therefore, characterizing classes of HAI scenarios in which this model structure works well, or doesn't, is another direction of future work.

6. CONCLUSIONS

6.1 Research Contributions

The contribution of this thesis is a probabilistic model of coupled human trust and self-confidence dynamics as they evolve during a human's interaction with automation. The dynamics are modeled as a partially observable Markov decision process that leverages behavioral and self-report data as observations for estimation of the cognitive states. Trust and self-confidence are modeled as separate discrete states whose transition probability functions are coupled. By doing so, the model is able to capture nuanced effects of various combinations of the states on the participant's reliance on the autonomous system. Moreover, the use of an asymmetrical structure in the emission probability functions that specifically captures the coupling of human reliance on automation to both trust and self-confidence enables labeling and interpretation of the coupled cognitive states. An experiment was designed and implemented to collect human behavioral and self-report data during their repeated interactions with an autonomous assistant in an obstacle avoidance game scenario. Using data collected from 340 human participants, the cognitive model was trained and validated. Analysis of the state transition probabilities suggests that participants' attribution of changes in performance to either themselves or the autonomous assistant vary depending on their states of trust and self-confidence. This underscores the importance of the proposed model for the design of human-aware automation, particularly in the context of human trust and self-confidence calibration in HAI. Future work includes validation of the model for other HAI scenarios, as well as model-based control algorithm design aimed at, for example, optimally allocating control authority to the human and autonomous system based on calibration of the human's cognitive states.

6.2 Future Work

There are several potential areas for future work to improve upon the model framework for increasing levels of application complexity. Examples of potential research directions include (1) introducing cognitive state-based feedback policies to properly allocate control authority,

(2) customizing the model framework to account for individuality, and (3) extending the proposed model framework to additional and increasingly complex contexts.

The model framework presented can be utilized to implement a model-based control algorithm design aimed at, for example, optimally allocating control authority to the human and autonomous system based, in part, on the human's cognitive states. This would enable the the automation to guide user behavior towards the best sequence of actions that maximize a specified performance objective or metric.

Work presented in this thesis assumes a single, general model for the population and does not account for individual demographic factors shown to influence dispositional trust, such as age, gender, and culture [16]. Future work may incorporate these individualistic factors within the model, or serve to identify behavioral clusters and customize group-specific models.

Finally, the model presented was developed and validated in the context of an online game-based task. Future work may aim to evaluate the robustness of the proposed model framework, specifically in evaluating whether the fundamental relationships between the actions, states, and observations are generalizable across contexts. Work may be done to extend the framework to more complex contexts, such as flight or driving simulators and real-world settings. Additionally, more sophisticated forms of automation assistance can be considered.

REFERENCES

[1] W. Barfield and T. A. Dingus, *Human Factors in Intelligent Transportation Systems*, 1st ed. New York: Taylor & Francis, 2014, ISBN: 978-1-317-78110-3.

[2] V. A. Banks, N. A. Stanton, and C. Harvey, "Sub-systems on the road to vehicle automation: Hands and feet free but not 'mind' free driving," en, *Safety Science*, vol. 62, pp. 505–514, Feb. 2014, ISSN: 0925-7535. DOI: 10.1016/j.ssci.2013.10.014.

[3] U. A. Force, "Report on Technology Horizons: A Vision for Air Force Science & Technology During 2010–203," en-US, Tech. Rep., 2010. [Online]. Available: https://www.airuniversity.af.edu/AUPress/Book-Reviews/Display/Article/1194559/report-on-technology-horizons-a-vision-for-air-force-science-technology-during/.

[4] U. E. Franke, "Drones, Drone Strikes, and US Policy: The Politics of Unmanned Aerial Vehicles," English, *Parameters*, vol. 44, no. 1, pp. 121–130, 2014, ISSN: 00311723.

[5] S. M. Astley, "Evaluation of computer-aided detection (CAD) prompting techniques for mammography," *The British Journal of Radiology*, vol. 78, no. suppl_1, S20–S25, Jan. 2005, ISSN: 0007-1285. DOI: 10.1259/bjr/37221979.

[6] E. de Visser, R. Pak, and T. Shaw, "From "automation" to "autonomy": The importance of trust repair in human-machine interaction," *Ergonomics*, vol. 61, pp. 1–33, Mar. 2018. DOI: 10.1080/00140139.2018.1457725.

[7] T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, en. Cambridge, MA: MIT Press, 1992, ISBN: 978-0-262-19316-0.

[8] M. Lind, "Plant modelling for human supervisory control," en, *Transactions of the Institute of Measurement and Control*, vol. 21, no. 4-5, pp. 171–180, Oct. 1999, ISSN: 0142-3312. DOI: 10.1177/014233129902100405.

[9] M. A. Goodrich and M. L. Cummings, "Human Factors Perspective on Next Generation Unmanned Aerial Systems," en, in *Handbook of Unmanned Aerial Vehicles*, K. P. Valavanis and G. J. Vachtsevanos, Eds., Dordrecht: Springer Netherlands, 2015, pp. 2405–2423, ISBN: 978-90-481-9707-1. DOI: 10.1007/978-90-481-9707-1_23.

[10] J. R. Peters, V. Srivastava, G. S. Taylor, A. Surana, M. P. Eckstein, and F. Bullo, "Human Supervisory Control of Robotic Teams: Integrating Cognitive Modeling with Engineering Design," *IEEE Control Systems Magazine*, vol. 35, no. 6, pp. 57–80, Dec. 2015, ISSN: 1941-000X. DOI: 10.1109/MCS.2015.2471056.

[11] A. Hussein, S. Elsawah, and H. Abbass, “Towards Trust-Aware Human-Automation Interaction: An Overview of the Potential of Computational Trust Models,” in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, Manoa, Hawaii, Jan. 2020. DOI: 10.24251/HICSS.2020.047.

[12] V. Riley, “Operator reliance on automation: Theory and data,” English, in *Automation and human performance: Theory and applications*, R. Parasuraman and M. Mouloua, Eds., 1st ed., Mahwah, NJ: CRC Press, Jun. 1996, pp. 19–35.

[13] J. D. Lee and N. Moray, “Trust, self-confidence, and operators’ adaptation to automation,” *International Journal of Human-Computer Studies*, vol. 40, no. 1, pp. 153–184, 1994. DOI: 10.1006/ijhc.1994.1007.

[14] J. D. Lee and K. A. See, “Trust in automation: Designing for appropriate reliance,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004. DOI: 10.1518/hfes.46.1.50_30392.

[15] J. Gao and J. D. Lee, “Extending the decision field theory to model operators’ reliance on automation in supervisory control situations,” *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 36, no. 5, pp. 943–959, Sep. 2006, ISSN: 1558-2426. DOI: 10.1109/TSMCA.2005.855783.

[16] K. A. Hoff and M. Bashir, “Trust in automation: Integrating empirical evidence on factors that influence trust,” *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 57, no. 3, pp. 407–434, 2015. DOI: 10.1177/0018720814547570.

[17] M. R. Endsley, “From Here to Autonomy: Lessons Learned From Human–Automation Research,” en, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 59, no. 1, pp. 5–27, Feb. 2017, ISSN: 0018-7208, 1547-8181. DOI: 10.1177/0018720816681350.

[18] P. A. Hancock, R. J. Jagacinski, R. Parasuraman, C. D. Wickens, G. F. Wilson, and D. B. Kaber, “Human-Automation Interaction Research: Past, Present, and Future,” *Ergonomics in Design*, vol. 21, no. 2, pp. 9–14, Apr. 2013, ISSN: 1064-8046. DOI: 10.1177/1064804613477099.

[19] R. Parasuraman and V. Riley, “Humans and Automation: Use, Misuse, Disuse, Abuse,” en, *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 39, no. 2, pp. 230–253, Jun. 1997, ISSN: 0018-7208, 1547-8181. DOI: 10.1518/001872097778543886.

[20] B. M. Muir, “Trust between humans and machines, and the design of decision aids,” *International Journal of Man-Machine Studies*, vol. 27, no. 5, pp. 527–539, 1987, ISSN: 0020-7373. DOI: https://doi.org/10.1016/S0020-7373(87)80013-5.

[21] J. G. Boubin, C. F. Rusnock, and J. M. Bindewald, "Quantifying Compliance and Reliance Trust Behaviors to Influence Trust in Human-Automation Teams," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 61, no. 1, pp. 750–754, Sep. 2017, ISSN: 2169-5067. DOI: 10.1177/1541931213601672.

[22] E. T. Chancey, J. P. Bliss, Y. Yamani, and H. A. H. Handley, "Trust and the Compliance–Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence," en, *Human Factors*, vol. 59, no. 3, pp. 333–345, May 2017, ISSN: 0018-7208. DOI: 10.1177/0018720816682648.

[23] O. Akbari and J. Sahibzada, "Students' Self-Confidence and Its Impacts on Their Learning Process," en, *American International Journal of Social Science Research*, vol. 5, no. 1, pp. 1–15, Jan. 2020, ISSN: 2576-1048, 2576-103X. DOI: 10.46281/aijssr.v5i1.462.

[24] J. D. Lee and N. Moray, "Trust, control strategies and allocation of function in human-machine systems," *Ergonomics*, vol. 35, no. 10, pp. 1243–1270, 1992. DOI: 10.1080/00140139208967392.

[25] J. Y. C. Chen and P. I. Terrence Peter I., "Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment.," *Ergonomics*, vol. 52, no. 8, pp. 907–920, Aug. 2009. DOI: 10.1080/00140130802680773.

[26] P. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," *International Journal of Human-Computer Studies*, Trust and Technology, vol. 58, no. 6, pp. 719–735, Jun. 2003, ISSN: 1071-5819. DOI: 10.1016/S1071-5819(03)00039-9.

[27] H. Neyedli, J. Hollands, and G. Jamieson, "Human Reliance on an Automated Combat ID System: Effects of Display Format," *Human Factors and Ergonomics Society Annual Meeting Proceedings*, vol. 53, pp. 212–216, Oct. 2009. DOI: 10.1518/107118109X12524441080.02.

[28] L. Wang, G. A. Jamieson, and J. G. Hollands, "The Effects of Design Features on Users' Trust in and Reliance on a Combat Identification System," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 55, no. 1, pp. 375–379, Sep. 2011, ISSN: 2169-5067. DOI: 10.1177/1071181311551077.

[29] P.-P. Maanen, F. Wisse, J. Diggelen, and R. J. Beun, "Effects of Reliance Support on Team Performance by Advising and Adaptive Autonomy," in *Proceedings of the 2011 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, Campus Scientifique de la Doua, Lyon, France, Aug. 2011, p. 287. DOI: 10.1109/WI-IAT.2011.117.

[30] D. Mikulski, F. Lewis, E. Gu, and G. Hudas, “Trust Method for Multi-Agent Consensus,” in *Proceedings of SPIE - The International Society for Optical Engineering*, Journal Abbreviation: Proceedings of SPIE - The International Society for Optical Engineering Publication Title: Proceedings of SPIE - The International Society for Optical Engineering, vol. 8387, Apr. 2012. DOI: 10.1117/12.918927.

[31] H. Saeidi and Y. Wang, “Trust and self-confidence based autonomy allocation for robotic systems,” in *2015 54th IEEE Conference on Decision and Control (CDC)*, Osaka, Japan.: IEEE, Dec. 2015, pp. 6052–6057. DOI: 10.1109/CDC.2015.7403171.

[32] I. Juvina, C. Lebiere, and C. Gonzalez, “Modeling trust dynamics in strategic interaction,” en, *Journal of Applied Research in Memory and Cognition*, Modeling and Aiding Intuition in Organizational Decision Making, vol. 4, no. 3, pp. 197–211, Sep. 2015, ISSN: 2211-3681. DOI: 10.1016/j.jarmac.2014.09.004.

[33] A. Xu and G. Dudek, “OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations,” in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI ’15, New York, NY, USA: ACM, 2015, pp. 221–228, ISBN: 978-1-4503-2883-8. DOI: 10.1145/2696454.2696492.

[34] M. W. Floyd, M. Drinkwater, and D. W. Aha, “Improving Trust-Guided Behavior Adaptation Using Operator Feedback,” en, in *Case-Based Reasoning Research and Development*, E. Hüllermeier and M. Minor, Eds., ser. Lecture Notes in Computer Science, Cham: Springer International Publishing, 2015, pp. 134–148, ISBN: 978-3-319-24586-7. DOI: 10.1007/978-3-319-24586-7_10.

[35] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, “Planning with Trust for Human-Robot Collaboration,” in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI ’18, New York, NY, USA: Association for Computing Machinery, Feb. 2018, pp. 307–315, ISBN: 978-1-4503-4953-6. DOI: 10.1145/3171221.3171264.

[36] B. Sadrfaridpour, M. F. Mahani, Z. Liao, and Y. Wang, “Trust-Based Impedance Control Strategy for Human-Robot Cooperative Manipulation,” American Society of Mechanical Engineers, Sep. 2018, V001T04A015–V001T04A015. DOI: 10.1115/DSCC2018-9170. [Online]. Available: https://proceedings.asmedigitalcollection.asme.org/proceeding.aspx?articleid=2715200.

[37] A. R. Wagner, P. Robinette, and A. Howard, “Modeling the Human-Robot Trust Phenomenon: A Conceptual Framework based on Risk,” *ACM Transactions on Interactive Intelligent Systems*, vol. 8, no. 4, 26:1–26:24, Nov. 2018, ISSN: 2160-6455. DOI: 10.1145/3152890.

[38] Y. Tao, E. Coltey, T. Wang, M. Alonso, M.-L. Shyu, S.-C. Chen, H. Alhaffar, A. Elias, B. Bogosian, and S. Vassigh, "Confidence Estimation Using Machine Learning in Immersive Learning Environments," in *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Shenzhen, Guangdong, China: IEEE, Aug. 2020, pp. 247–252, ISBN: 978-1-72814-272-2. DOI: 10.1109/MIPR49039.2020.00058.

[39] H. Azevedo-Sa, S. K. Jayaraman, C. T. Esterwood, X. J. Yang, L. P. Robert, and D. M. Tilbury, "Real-Time Estimation of Drivers' Trust in Automated Driving Systems," en, *International Journal of Social Robotics*, Sep. 2020, ISSN: 1875-4805. DOI: 10.1007/s12369-020-00694-1.

[40] H. Saeidi and Y. Wang, "Incorporating Trust and Self-Confidence Analysis in the Guidance and Control of (Semi)Autonomous Mobile Robotic Systems," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 239–246, Apr. 2019, ISSN: 2377-3766. DOI: 10.1109/LRA.2018.2886406.

[41] M. R. A. R. Rani, M. A. Sinclair, and K. Case, "Human mismatches and preferences for automation," *International Journal of Production Research*, vol. 38, no. 17, pp. 4033–4039, Nov. 2000, ISSN: 0020-7543. DOI: 10.1080/002075400050204894.

[42] R. Wiczorek and J. Meyer, "Effects of Trust, Self-Confidence, and Feedback on the Use of Decision Automation," English, *Frontiers in Psychology*, vol. 10, p. 519, 2019, ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.00519.

[43] O. Sigaud and O. Buffet, *Markov Decision Processes in Artificial Intelligence*, ser. ISTE. Hoboken, New Jersey: John Wiley & Sons, 2013, ISBN: 978-1-118-62010-6.

[44] K. Akash, G. McMahon, T. Reid, and N. Jain, "Human Trust-Based Feedback Control: Dynamically Varying Automation Transparency to Optimize Human-Machine Interactions," *IEEE Control Systems Magazine*, vol. 40, no. 6, pp. 98–116, Dec. 2020, ISSN: 1941-000X. DOI: 10.1109/MCS.2020.3019151.

[45] *Amazon Mechanical Turk*. [Online]. Available: https://www.mturk.com/.

[46] Ó. Pérez, M. Piccardi, J. García, M. Á. Patricio, and J. M. Molina, "Comparison Between Genetic Algorithms and the Baum-Welch Algorithm in Learning HMMs for Human Activity Classification," en, in *Applications of Evolutionary Computing*, ser. Lecture Notes in Computer Science, M. Giacobini, Ed., vol. 4448, Berlin, Heidelberg: Springer, 2007, pp. 399–406, ISBN: 978-3-540-71805-5.

[47] *MATLAB Optimization Toolbox*, Natick, MA, 2021. [Online]. Available: https://www.mathworks.com/products/optimization.html#resources.

[48] L. Rabiner and B.-H. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986. DOI: 10.1109/MASSP.1986.1165342.

[49] B. M. Muir, "Operators' trust in and use of automatic controllers in a supervisory process control task.," en, Ph.D. Thesis, University of Toronto, Toronto, ON, Canada, 1990.

[50] S. Lewandowsky, M. Mundy, and G. P. A. Tan, "The dynamics of trust: Comparing humans to automation," *Journal of Experimental Psychology: Applied*, vol. 6, no. 2, pp. 104–123, 2000, ISSN: 1939-2192(Electronic),1076-898X(Print). DOI: 10.1037/1076-898X.6.2.104.

[51] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, Mar. 1996, ISSN: 0014-0139. DOI: 10.1080/00140139608964474.

[52] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," en, *International Journal of Human-Computer Studies*, Trust and Technology, vol. 58, no. 6, pp. 697–718, Jun. 2003, ISSN: 1071-5819. DOI: 10.1016/S1071-5819(03)00038-7.

[53] B. Weiner, *An Attribution Theory of Motivation and Emotion*. New York, NY: Springer-Verlag, Jan. 1986, vol. 92, ISBN: 978-1-4612-9370-5.

[54] K. van Dongen and P.-P. van Maanen, "A framework for explaining reliance on decision aids," en, *International Journal of Human-Computer Studies*, vol. 71, no. 4, pp. 410–424, Apr. 2013, ISSN: 1071-5819. DOI: 10.1016/j.ijhcs.2012.10.018.

A. TRAINED MODEL RESULTS

Here I present the POMDP model of human trust–self-confidence behavior discussed in Chapter 4.

A.1 Initial State Probabilities

The initial state probabilities for trust $\pi_T : 1 \times T \rightarrow [0, 1]$ and self-confidence π_{SC} : $1 \times SC \rightarrow [0, 1]$ are both represented by 1×2 matrices that represent the probability of the initial trust state s_T and self-confidence state s_{SC} respectively. The initial state probabilities are provided in Table A.1.

Table A.1. Initial trust state s_T and self-confidence state s_{SC} probabilities

Trust		Self-Confidence	
T↓	**T↑**	**SC↓**	**SC↑**
0.1878	0.8122	0.6070	0.3930

A.2 Transition Probabilities

The transition probabilities for trust $\mathcal{T}_T : \mathcal{S} \times T \times \mathcal{A} \rightarrow [0, 1]$ and self-confidence $\mathcal{T}_{SC}$: $\mathcal{S} \times SC \times \mathcal{A} \rightarrow [0, 1]$ are each represented by $4 \times 2 \times 18$ matrices that map the probability of transitioning from combinations of states $\mathcal{S}$ of trust $s_T \in T$ and self-confidence $s_{SC} \in SC$ to the next states of trust and self-confidence, respectively, given an action $a \in \mathcal{A}$. The state combination transition probabilities are the product of the individual transition probabilities of trust and self-confidence, as given by

$$\mathcal{T}(s'|s, a) = \mathcal{T}(s'_T|s_T, s_{SC}, a)\mathcal{T}(s'_{SC}|s_T, s_{SC}, a) \quad . \tag{A.1}$$

The transition probabilities are provided in Tables A.2- A.4. The transition probability tables are separated by the action a_A. Each table is divided such that the transition probabilities can be identified based upon the change in performance metrics.

Table A.2. Transition probabilities for $a_A \in \Theta_L$ and performance metric combinations

Collision Decrease, Time Decrease					**Collision Decrease, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓'	**T↑'**	**SC↓'**	**SC↑'**		**T↓'**	**T↑'**	**SC↓'**	**SC↑'**
T↓SC↓	0.0019	0.9981	0.9992	0.0008	**T↓SC↓**	0.9959	0.0041	0.8142	0.1858
T↓SC↑	0.9990	0.0010	0.0037	0.9963	**T↓SC↑**	0.8518	0.1482	0.0003	0.9997
T↑SC↓	0.7308	0.2692	0.8219	0.1781	**T↑SC↓**	0.0011	0.9989	0.9696	0.0304
T↑SC↑	0.0403	0.9597	0.0298	0.9702	**T↑SC↑**	0.0001	0.9999	0.0158	0.9842
Collision No Change, Time Decrease					**Collision No Change, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓'	**T↑'**	**SC↓'**	**SC↑'**		**T↓'**	**T↑'**	**SC↓'**	**SC↑'**
T↓SC↓	0.4572	0.5428	0.7949	0.2051	**T↓SC↓**	0.0002	0.9998	0.9512	0.0488
T↓SC↑	0.9738	0.0262	0.0030	0.9970	**T↓SC↑**	0.9534	0.0466	0.0635	0.9365
T↑SC↓	0.9997	0.0003	0.9612	0.0388	**T↑SC↓**	0.0296	0.9704	0.9999	0.0001
T↑SC↑	0.0013	0.9987	0.0074	0.9926	**T↑SC↑**	0.0074	0.9926	0.0266	0.9734
Collision Increase, Time Decrease					**Collision Increase, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓'	**T↑'**	**SC↓'**	**SC↑'**		**T↓'**	**T↑'**	**SC↓'**	**SC↑'**
T↓SC↓	0.9990	0.0010	0.9552	0.0448	**T↓SC↓**	0.0022	0.9978	0.9992	0.0008
T↓SC↑	0.9982	0.0018	0.1574	0.8426	**T↓SC↑**	0.9965	0.0035	0.0054	0.9946
T↑SC↓	0.0844	0.9156	0.9960	0.0040	**T↑SC↓**	0.0004	0.9996	0.9988	0.0012
T↑SC↑	0.4409	0.5591	0.1010	0.8990	**T↑SC↑**	0.6822	0.3178	0.0133	0.9867

A.3 Emission Probabilities

The emission probability function for reliance $\mathcal{E}_R : \mathcal{S} \times R \rightarrow [0,1]$ is represented by a 4×2 matrix that maps the probability of reliance on automation $o_R \in R$ given the current trust and self-confidence belief states. The emission probability function for self-reported self-confidence $\mathcal{E}_{srSC} : SC \times srSC \rightarrow [0,1]$ is represented by a 2×2 matrix that maps

Table A.3. Transition probabilities for $a_A \in \Theta_M$ and performance metric combinations

Collision Decrease, Time Decrease					**Collision Decrease, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓’	**T↑’**	**SC↓’**	**SC↑’**		**T↓’**	**T↑’**	**SC↓’**	**SC↑’**
T↓SC↓	0.9838	0.0162	0.9963	0.0037	**T↓SC↓**	0.9940	0.0060	0.9919	0.0081
T↓SC↑	0.9973	0.0027	0.0021	0.9979	**T↓SC↑**	0.9232	0.0768	0.0019	0.9981
T↑SC↓	0.7759	0.2241	0.6608	0.3392	**T↑SC↓**	0.1517	0.8483	0.7768	0.2232
T↑SC↑	0.0621	0.9379	0.0034	0.9966	**T↑SC↑**	0.0753	0.9247	0.0293	0.9707
Collision No Change, Time Decrease					**Collision No Change, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓’	**T↑’**	**SC↓’**	**SC↑’**		**T↓’**	**T↑’**	**SC↓’**	**SC↑’**
T↓SC↓	0.9788	0.0212	0.9720	0.0280	**T↓SC↓**	0.0030	0.9970	0.9976	0.0024
T↓SC↑	0.9922	0.0078	0.0015	0.9985	**T↓SC↑**	0.9983	0.0017	0.0033	0.9967
T↑SC↓	0.5040	0.4960	0.8412	0.1588	**T↑SC↓**	0.0018	0.9982	0.9599	0.0401
T↑SC↑	0.0323	0.9677	0.0230	0.9770	**T↑SC↑**	0.0000	1.0000	0.0462	0.9538
Collision Increase, Time Decrease					**Collision Increase, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓’	**T↑’**	**SC↓’**	**SC↑’**		**T↓’**	**T↑’**	**SC↓’**	**SC↑’**
T↓SC↓	0.9989	0.0011	0.9998	0.0002	**T↓SC↓**	0.0013	0.9987	0.9984	0.0016
T↓SC↑	0.9740	0.0260	0.1244	0.8756	**T↓SC↑**	0.9933	0.0067	0.1167	0.8833
T↑SC↓	0.7311	0.2689	0.9735	0.0265	**T↑SC↓**	0.9959	0.0041	0.9084	0.0916
T↑SC↑	0.0531	0.9469	0.1092	0.8908	**T↑SC↑**	0.0601	0.9399	0.1858	0.8142

the probability of low or high self-reported self-confidence $o_{srSC} \in srSC$ given the current self-confidence state. The overall emission probabilities are the product of the reliance and self-reported self-confidence emission probabilities, given by

$$\mathcal{E}(o|s) = \mathcal{E}(o_R|s_T, s_{SC})\mathcal{E}(o_{srSC}|s_{SC}) \ . \tag{A.2}$$

The emission probabilities are provided in Table A.5.

Table A.4. Transition probabilities for $a_A \in \Theta_H$ and performance metric combinations

Collision Decrease, Time Decrease					**Collision Decrease, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓’	**T↑’**	**SC↓’**	**SC↑’**		**T↓’**	**T↑’**	**SC↓’**	**SC↑’**
T↓SC↓	0.4471	0.5529	0.7465	0.2535	**T↓SC↓**	0.4109	0.5891	0.6672	0.3328
T↓SC↑	1.0000	0.0000	0.0014	0.9986	**T↓SC↑**	0.9199	0.0801	0.0011	0.9989
T↑SC↓	0.9935	0.0065	0.9992	0.0008	**T↑SC↓**	0.0005	0.9995	1.0000	0.0000
T↑SC↑	0.0027	0.9973	0.0487	0.9513	**T↑SC↑**	0.0382	0.9618	0.0048	0.9952
Collision No Change, Time Decrease					**Collision No Change, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓’	**T↑’**	**SC↓’**	**SC↑’**		**T↓’**	**T↑’**	**SC↓’**	**SC↑’**
T↓SC↓	0.7048	0.2952	0.9986	0.0014	**T↓SC↓**	0.0010	0.9990	1.0000	0.0000
T↓SC↑	0.9958	0.0042	0.0013	0.9987	**T↓SC↑**	0.9453	0.0547	0.0061	0.9939
T↑SC↓	0.0071	0.9929	0.8432	0.1568	**T↑SC↓**	0.5768	0.4232	0.8019	0.1981
T↑SC↑	0.0003	0.9997	0.0012	0.9988	**T↑SC↑**	0.0127	0.9873	0.0021	0.9979
Collision Increase, Time Decrease					**Collision Increase, Time Increase**				
	Trust		Self-Confidence			Trust		Self-Confidence	
	T↓’	**T↑’**	**SC↓’**	**SC↑’**		**T↓’**	**T↑’**	**SC↓’**	**SC↑’**
T↓SC↓	0.0020	0.9980	0.9677	0.0323	**T↓SC↓**	0.0160	0.9840	0.9998	0.0002
T↓SC↑	0.9923	0.0077	0.1525	0.8475	**T↓SC↑**	0.9853	0.0147	0.0015	0.9985
T↑SC↓	0.8208	0.1792	0.9524	0.0476	**T↑SC↓**	0.9973	0.0027	0.9979	0.0021
T↑SC↑	0.0683	0.9317	0.0828	0.9172	**T↑SC↑**	0.0350	0.9650	0.0456	0.9544

Table A.5. Emission probabilities of the reliance observation o_R and self-reported self-confidence observation o_{srSC}. NR and R denote no reliance and reliance respectively, while high and low self-reported self-confidence is denoted by $srSC\uparrow$ and $srSC\downarrow$ respectively.

	Reliance		Self-Reported Self-Confidence		
	NR	**R**		**srSC↓**	**srSC↑**
T↓SC↓	0.4862	0.5138	**SC↓**	0.9448	0.0552
T↓SC↑	0.8954	0.1046	**SC↑**	0.0898	0.9102
T↑SC↓	0.4983	0.5017			
T↑SC↑	0.1083	0.8917			