# A Computational Model of Coupled Human Trust and Self-confidence Dynamics

KATHERINE J. WILLIAMS, MADELEINE S. YUH, and NEERA JAIN, Purdue University, USA

Autonomous systems that can assist humans with increasingly complex tasks are becoming ubiquitous. Moreover, it has been established that a human's decision to rely on such systems is a function of both their trust in the system and their own self-confidence as it relates to executing the task of interest. Given that both under- and over-reliance on automation can pose significant risks to humans, there is motivation for developing autonomous systems that could appropriately calibrate a human's trust or self-confidence to achieve proper reliance behavior. In this article, a computational model of coupled human trust and self-confidence dynamics is proposed. The dynamics are modeled as a partially observable Markov decision process without a reward function (POMDP/R) that leverages behavioral and self-report data as observations for estimation of these cognitive states. The model is trained and validated using data collected from 340 participants. Analysis of the transition probabilities shows that the proposed model captures the probabilistic relationship between trust, self-confidence, and reliance for all discrete combinations of high and low trust and self-confidence. The use of the proposed model to design an optimal policy to facilitate trust and self-confidence calibration is a goal of future work.

CCS Concepts: • **Human-centered computing → HCI theory, concepts and models**; **Empirical studies in HCI**;

Additional Key Words and Phrases: Human cognitive modeling, human trust in automation, human self-confidence, computational modeling, partially observable Markov decision process

**ACM Reference format:**
Katherine J. Williams, Madeleine S. Yuh, and Neera Jain. 2023. A Computational Model of Coupled Human Trust and Self-confidence Dynamics. *ACM Trans. Hum.-Robot Interact.* 12, 3, Article 39 (June 2023), 29 pages.
https://doi.org/10.1145/3594715



## 1 INTRODUCTION

The complexity of human interactions with autonomous systems is increasing, as evidenced in applications including intelligent transportation systems [11], autonomous vehicles [10], military operations [28, 29], and medical imaging systems [8]. In turn, this necessitates a greater understanding of these interactions and how they affect outcomes in terms of metrics such as

K. J. Williams and M. S. Yuh contributed equally to this research.
This material is based upon work supported by the National Science Foundation under Award No. CNS-1836952. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.
Authors' address: K. J. Williams, M. S. Yuh, and N. Jain, Purdue University, 177 S. Russell St., West Lafayette, IN 47907-2099; emails: Katherine.Williams.me@gmail.com, {myuh, neerajain}@purdue.edu.

2573-9522/2023/06-ART39 $15.00
https://doi.org/10.1145/3594715

performance [19, 32, 46, 67]. 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)** [38, 57]. In particular, the cognitive factors of human trust and self-confidence play a substantial role in the human's willingness, and decision, to rely on automation [25, 31, 33, 34, 43, 44, 62]. Interestingly, the inclusion of automation support presents the potential for automation bias—an over-reliance on an automated decision aid, in which the human attributes more authority to the automation than to other sources [55]. This often results in the human neglecting prior knowledge and contradictory evidence to follow incorrect advice. Consequences of improper reliance, relying too much or too little, can be dire [56]. For example, it is well supported that miscalibration of trust to automation capabilities is the cause of misuses and disuse of automation [44]. This motivates the need for *calibration* of cognitive factors to achieve appropriate reliance. For example, models enabling cognitive state estimation and prediction could be used by automation to appropriately trigger system responses through methods such as transparency adaptation, automation behavior adaptation, or flexible autonomy [5, 38]. However, accomplishing this often 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 different cognitive factors in human behavior and decision-making, particularly as it relates to human *reliance* on automation [12, 15, 21, 25, 26, 34, 44, 48, 50, 56, 71]. A majority of these frameworks are centered around human *trust* in automation [12, 15, 25, 34] and its effect on reliance. Trust is well established as a cognitive factor that can be defined in an HAI context as the belief that the automation will help the human achieve their goal(s) in an uncertain situation [44]. Early qualitative models of trust establish that human trust in automation is dependent on factors including interaction with the operator, context, automation performance, and the user interface [44]. Another widely referenced qualitative model by Hoff and Bashir [34] identifies three stratified layers of trust: dispositional trust—derived from individual characteristics and remains characteristically constant over time; situational trust—derived from the environment; and learned trust—derived from preexisting knowledge and the system's performance. In turn, researchers have highlighted factors that affect the human's trust, including system transparency [77], anthropomorphism [18], and automation reliability [13, 20]. However, in addition to trust, it has been established that the *self-confidence* of the human also affects their reliance on automation [16, 22, 42, 43, 53, 56, 73]. For example, over-reliance on automation can arise as a result of a human with low self-confidence in their skill to manually execute a particular task [56]. Additionally, biases in one's self-confidence (over- or under-confidence) can lead to improper reliance [43].

There has been a significant effort over the last decade to develop computational models for predicting reliance behavior or the dynamics of trust and self-confidence. An overview of computational models of human trust or self-confidence is provided in Table 1. From a computational perspective, several models have been developed to predict human trust, particularly in the last decade. Notably, more recent models of trust are aimed at capturing the probabilistic nature of human behavior using a variety of mathematical techniques. Computational cognitive models of trust include **auto-regressive moving average vector (ARMAV)** derivations and other linear models [9, 35, 37, 42], decision analytical models based on decision or game theory [76], dynamic Bayesian networks [27, 30, 72], and **partially observable Markov decision process (POMDP)** models [5, 6, 17]. However, despite several conceptual frameworks supporting the relationship between trust and self-confidence, comparatively fewer *computational* models have been developed to capture this relationship [31, 43, 65]. Many of these models are based upon the *"confidence vs. trust" hypothesis*, originally developed in [43], that assumes a human's reliance on a given system is dependent on a difference between the human's trust in the automation and confidence

Table 1. Summary of Models of Trust or Self-confidence

| Papers | Category | | | |
|---|---|---|---|---|
| | Trust | Self-confidence | T-SC Coupling | Probabilistic |
| Lee and Moray, 1992 [42] | ✓ | | | |
| Lee and Moray, 1994* [43] | ✓ | ✓ | | |
| Gao and Lee, 2006* [31] | ✓ | ✓ | | ✓ |
| Maanen et al., 2011 [47] | ✓ | | | |
| Mikulski et al., 2012 [49] | ✓ | | | ✓ |
| Saeidi and Wang, 2015* [64] | ✓ | ✓ | | |
| Juvina et al., 2015 [40] | ✓ | | | ✓ |
| Xu and Dudek, 2015 [76] | ✓ | | | ✓ |
| Floyd et al., 2015 [27] | ✓ | | | |
| Hu et al., 2016 [36] | ✓ | | | |
| Akash et al., 2017 [7] | ✓ | | | |
| Akash et al., 2018 [4] | ✓ | | | |
| DeVisser et al., 2018 [19] | ✓ | | | |
| Chen et al., 2018 [17] | ✓ | | | ✓ |
| Sadrfaridpour et al., 2018 [63] | ✓ | | | ✓ |
| Wagner et al., 2018 [72] | ✓ | | | |
| Hu et al., 2019 [37] | ✓ | | | |
| Juvina et al., 2019 [39] | ✓ | | | |
| Saeidi and Wang, 2019 [65] | ✓ | ✓ | | ✓ |
| Tao et al., 2020 [70] | | ✓ | | ✓ |
| Akash et al., 2020 [5] | ✓ | | | ✓ |
| Azevedo-Sa et al., 2020 [9] | ✓ | | | |
| Soh et al., 2020 [69] | ✓ | | | ✓ |

"T-SC Coupling" refers to models that capture the relationship between trust and self-confidence while recognizing that these two individual states affect one another dynamically. *denotes models that are based upon the *"confidence vs. trust" hypothesis.*

in their ability (also known as the *relative trust*) to execute the task manually [71]. 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 [61, 75]. For example, in [75], 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 [61] 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.* This "coupling" refers to models that capture the relationship between trust and self-confidence while *also* recognizing that these two individual states affect one another *dynamically*. While prior work has explored the coupled relationship between trust and workload [3], to the knowledge of the authors, there are no existing models that mathematically characterize the dynamic coupling between trust and self-confidence.

The primary contribution of this article is a probabilistic discrete-state model of human trust and self-confidence dynamics as they relate to a human's repeated interactions with automation assistance. 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 automation 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 [61, 75] 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 article is organized as follows. In Section 2, existing computational models of trust and self-confidence are presented and discussed in greater detail, along with a comparison to the proposed approach. In Section 3, the formulation of the trust and self-confidence modeling framework is presented. The human subject study, including experimental design and implementation, is outlined in Section 4. The modeling, training, and validation process is discussed in Section 5. The trained model is analyzed in Section 6, followed by a discussion of the implications of the results on the design of human-responsive automation and limitations of the work. Finally, conclusions and future research directions are discussed in Section 7.

## 2 RELATED WORK

Before presenting our modeling approach, we describe in greater detail the existing computational models that relate human trust and self-confidence in HAI contexts. A review of applied quantitative models of trust is available in [66]. Lee and Moray [42] first developed an ARMAV time series model in 1992 to model trust as a function of performance efficiency and system faults. This model was extended in 1994 to capture the relationship between trust, self-confidence, and reliance on automation, after identifying that the use of automatic control was strongly correlated to the difference between users' trust and self-confidence. This model is provided in Equation (1) and predicts the operators' allocation strategy by means of the percentage of automatic control [43]. The model accounts for past automation dependence, a difference in the operators' trust and self-confidence states, as well as individual operator bias. The variable $\phi$ is a constant representing the current use of automation dependence on past use of automation. The variables $A1$ and $A2$ represent the weights of the difference in trust and self-confidence $(T - SC)$ and individual bias toward manual operation, respectively. Normally distributed independent fluctuations are provided by $a(t)$, given time $t$.

$$\%Automatic = \phi 1 \times Automatic(t-1) + A1 \times ((T - SC)(t)) + A2 \times IndividualBias + a(t) \quad (1)$$

Gao and Lee [31] developed an alternative model that utilizes the difference between trust and self-confidence to determine reliance on automation behavior like that of Lee and Moray [43]. The *EDFT* model is an extended **decision field theory (DFT)** model [14] used to characterize multiple decisions made sequentially, as opposed to the single decisions addressed by DFT. The EDFT model structure utilizes a closed-loop relationship between the context (autonomous $C_A$ and manual $C_M$), information available, operator belief (context autonomous $B_{CA}$ and manual $B_{CM}$), cognitive state (trust $T$ and self-confidence $SC$), intention ($P$), and decision (reliance). The preference, $PR$, of mode is defined as the difference between trust and self-confidence (Equation (2)) and updated given the context and noise term $\epsilon$ representing the uncertainty in trust or self-confidence in Equation (3). The model is then used to predict the user's decision to rely on automation or to use manual control when the preference evolves beyond a given threshold $\theta$.

$$PR(n) = T(n) - SC(n) \quad (2)$$

$$PR(n) = (1 - s) \times PR(n-1) + s \times [C_A(n-1) - C_M(n-1)] + \epsilon(n) \tag{3}$$

In 2015, Saeidi and Wang [64] developed a performance-based, computational trust and self-confidence model, *TSC* (Equation (4)), for autonomy allocation in a UAV context. The TSC model is a function of the human's ($P_h$) and robot's ($P_r$) performance at time step $k$, with performance level constants $a_T$ and $b_T$. Similar to the models of [31, 43], this model incorporates a *difference* in the cognitive states, human-to-robot trust and self-confidence, to achieve optimal allocation with consideration of the Yerkes-Dodson law [23] and robot performance decay. Therefore, to reduce the effects of human workload overload or poor robot performance, the level of autonomy is switched to maintain the difference, TSC, within the given thresholds. This difference is depicted in Equation (4).

$$TSC(k) = a_T P_r(k) - b_T P_H(k) \tag{4}$$

In 2019, Saeidi and Wang [65] improved upon their trust and self-confidence allocation strategy by incorporating a TSC-based switching control for manual and fully autonomous mode allocation. They model the difference between trust and self-confidence as a direct function of human and robot performance. This is similar to how Lee and Moray [42] model trust as a function of performance efficiency. Lee and Moray [43] denote T-SC as the difference between subjective ratings of trust and self-confidence, and Gao and Lee [31] treat trust and self-confidence as a function of the operator's belief in the automation and manual control capability. The proposed model in this article will be considering task performance as an action, or input, that affects the dynamic evolution of the cognitive states of trust and self-confidence, along with other environmental and task context factors that will be expanded upon in Section 3. Furthermore, among these existing computational models incorporating cognitive states of both trust and self-confidence, there is a key similarity in the basis of their frameworks. This similarity is the idea of the *"confidence vs. trust" hypothesis*, or assuming the human's reliance on a given system is dependent on a *difference* between the human's trust in the automation and confidence in their individual ability. On the other hand, by incorporating cognitive state coupling of the human's trust and self-confidence, the model proposed in this article is unique in its ability to capture the relationship between trust and self-confidence while recognizing that these two individual states affect one another dynamically.

## 3 MODEL DEFINITION

A POMDP is an extension of a **Markov decision process (MDP)** and is defined as a 7-tuple, $(\mathcal{S}, \mathcal{A}, O, \mathcal{T}, \mathcal{E}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, and $O$ is a finite set of observations [68]. 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 $\gamma$ can be used to synthesize an optimal action (control) policy given the state dynamics. However, designing such a policy is outside the scope of this work; therefore, throughout the remainder of the article, we will refer to the 5-tuple $(\mathcal{S}, \mathcal{A}, O, \mathcal{T}, \mathcal{E})$ as a POMDP/R.

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. The POMDP is used 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 described in more detail below. It is worth noting that POMDPs are often used in robotic contexts in which the states are the robot's current position, the actions are the

possible directions the robot can travel in, and the observation is the robot's future position [58]. However, *here* we model a human's cognitive behavior using a POMDP, as is done in [5]. To do so, we define relevant human cognitive factors as the states of the POMDP, actions are the measures that influence the cognitive states (namely characteristics of the automation's input as well as the human's experience with it), and observations are the observable characteristics of the human's decision.

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 dynamics and was shown to be sufficient for real-time trust calibration [5]. 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 the task environment) as well as the human's experience with the automation. The latter is characterized here as the system performance, which reflects the calculated score earned by the participant in the previous trial. For example, a participant's trust is a function of their performance in the previous trial. This means that transitions in the trust state are driven by the change in performance between the previous and current trials. It should be noted that because the model states are defined as factors of the human's cognition, both uncontrollable and controllable actions affect the state dynamics [5]. The *Automation Input* $a_A$ from the agent is controllable and belongs to the controllable action set $\mathcal{A}_c$. However, the system *Performance* $a_P$ is considered uncontrollable from the agent's perspective as it is driven in part by the human's behavior, and therefore belongs to the uncontrollable action set $\mathcal{A}_{uc}$. In other words, the POMDP/R in this article is a 6-tuple, $(\mathcal{S}, \mathcal{A}_{uc}, \mathcal{A}_c, \mathcal{O}, \mathcal{T}, \mathcal{E})$. Nevertheless, for consistency with the standard definition of a POMDP/R, we will combine the controllable and uncontrollable action sets into one action set such that $\mathcal{A} = \{\mathcal{A}_{uc}, \mathcal{A}_c\}$. Supported by the literature discussed in Section 1 citing the coupling between human trust and self-confidence, the states are assumed to be coupled according to the following transition probability functions: $\mathcal{T}(s'_T|s_T, s_{SC}, a)$ and $\mathcal{T}(s'_{SC}|s_T, s_{SC}, a)$.

Finally, the set of observations $\mathcal{O}$ is defined as the observable characteristics of the human's behavior and decision-making. As discussed earlier, it is well established in the literature that human reliance on automation is affected by both the human's trust in the automation and their self-confidence [24, 44, 56]. In other words, reliance is specifically defined as an observation (as opposed to an action) in the POMDP/R, with the emission probability function for reliance defined as $\mathcal{E}(o_R|s_T, s_{SC})$. It is worth noting that although a user's past reliance decision could be construed as a predictor of their future trust in the automation, it is their performance resulting from a reliance decision that actually influences their state of trust. This further underscores the choice of reliance as an observation and performance (as a proxy of experience with the automation) as an uncontrollable action.

While a POMDP/R 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 [70]. This creates asymmetry in the emission probability function that aids interpretability of the model, as discussed in Section 5. The proposed POMDP model definition is summarized in Table 2 and depicted in Figure 1. For ease of notation, we will denote uncontrollable actions $\mathcal{A}_{uc}$ as $A_p$ such that $a_{uc,P} = a_P$ and controllable actions $\mathcal{A}_c$ as $A_A$ such that $a_{c,A} = a_A$ going forward.

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 (5), in which $P(\cdot)$

Table 2. Definition of the Human Trust–Self-confidence (T-SC) POMDP/R Model

| | | |
|---|---|---|
| States $s \in \mathcal{S}$ | $\mathcal{S} = \begin{bmatrix} \text{Trust } s_T \\ \text{Self-confidence } s_{SC} \end{bmatrix}$ | $s_T \in T$<br>$T = \begin{Bmatrix} \text{Low Trust } T\downarrow \\ \text{High Trust } T\uparrow \end{Bmatrix}$<br>$s_{SC} \in SC$<br>$SC = \begin{Bmatrix} \text{Low Self-confidence } SC\downarrow \\ \text{High Self-confidence } SC\uparrow \end{Bmatrix}$ |
| Actions $a \in \mathcal{A}$ | $\mathcal{A} = \{\mathcal{A}_c, \mathcal{A}_{uc}\}$<br>$\mathcal{A}_{uc}$ := Performance $a_{uc,P}$<br>$\mathcal{A}_c$ := Automation Input $a_{c,A}$ | $a_{uc,P} \in \mathcal{A}_{uc}$<br>$\mathcal{A}_{uc} = \begin{Bmatrix} \text{Performance Deterioration } P^- \\ \text{Performance Improvement } P^+ \end{Bmatrix}$<br>$a_{c,A} \in \mathcal{A}_c$<br>$\mathcal{A}_c$ : Context Specific |
| Observations $o \in O$ | $O = \begin{bmatrix} \text{Reliance } o_R \\ \text{Self-reported Self-Confidence } o_{srSC} \end{bmatrix}$ | $o_R \in R$<br>$R = \begin{Bmatrix} \text{No Reliance } R_{NR} \\ \text{Reliance } R_R \end{Bmatrix}$<br>$o_{srSC} \in srSC$<br>$srSC = \begin{Bmatrix} \text{Low Self-confidence } srSC\downarrow \\ \text{High Self-confidence } srSC\uparrow \end{Bmatrix}$ |

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/R.

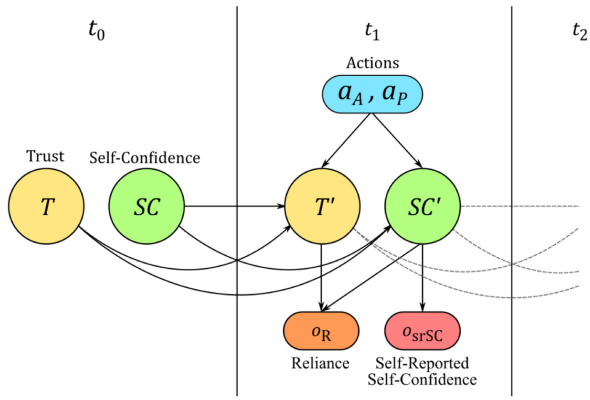



Fig. 1. A representation of the proposed POMDP/R model of trust and self-confidence. The transition probabilities of trust and self-confidence depend on the previous states of trust and self-confidence. The reliance observation is dependent on both the trust state and self-confidence state. However, the self-reported self-confidence observation is dependent on only the self-confidence state.

denotes probability.

$$b'(s') = P(s'|o, a, b(s)) = \frac{P(o|s', a) \sum_{s \in S} P(s'|s, a) b(s)}{\sum_{s' \in S} P(o|s', a) \sum_{s \in S} P(s'|s, a) b(s)} \tag{5}$$

## 4 HUMAN SUBJECT STUDY

In Section 4.1, 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 4.2, and analysis of behavioral and self-report data collected from the experiment is presented in Section 4.3.

### 4.1 Study Design

Human subject data is collected in the context of a 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 six obstacles. The participants are also informed that an automation assistant is available to help them play the game. Note that in reality, the "automation assistant" simply scales the user's mouse input by a pre-assigned parameter $\theta$. The scaling factor $\theta$ can take on values belonging to any one of three sets: $\Theta_L = \{0.7, 0.8, 0.9\}$, $\Theta_M = \{1.0, 1.1, 1.2\}$, and $\Theta_H = \{1.3, 1.4, 1.5\}$, where $\theta \in \Theta_j$ for $j = \{L, M, H\}$. 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 automation assistance algorithm, the value of $\theta$ experienced by each participant is assigned to them according to the between-subjects study design described below. In other words, the scaling factor is not responsive to the human's performance. Rather, the goal of the experiment is to obtain a set of training data that captures the effect of a range of values of the automation assistant's input on participants' behavior. Whether a particular value of $\theta$ helps or hinders the participant is a function of their skill level. For example, automation input values belonging to $\Theta_L$ scale the user's input down. While this may be beneficial for a user whose mouse input is over-reacting, the assistance may not help a user who is already playing well. This is by design to stimulate changes in the user's trust and, in turn, reliance on the automation assistance. For example, we expect that a user whose performance is being aided by the automation assistance will choose to continue to rely on it, whereas a user who finds the automation to be inhibiting their performance will do the opposite. Stimulating both increases and decreases in trust is critical for collecting training data that covers the state space of interest—in this case, all discrete combinations of low and high trust and self-confidence.

In the game shown in Figure 2, the penguin avatar moves at a constant speed, and its position is controlled by the participant's mouse movement. The penguin's $x$ and $y$ positions are governed by the following dynamical equations:

$$\begin{aligned} x_{t+1} &= x_t + \Delta t V cos(\theta_k u_t) + \phi(y) \\ y_{t+1} &= y_t + \Delta t V sin(\theta_k u_t), \end{aligned} \tag{6}$$

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 $k = 1, \ldots, 10$. During the practice round, participants do not receive any input scaling, so $\theta_0 = 1$. The game update discrete time interval is $\Delta t$, $V$ is the constant speed, and $\phi(y)$ is an added "wind" effect that increases in the upward vertical direction and is defined relative to the maximum vertical position, $y_{max}$. Table 3 provides the specific parameter values used in the experiment. It is important to note that the automation never takes control away from the participant.

A between-subjects study is designed to elicit changes in each participant's trust in the automation assistant *and* confidence in their ability to play the game (i.e., their self-confidence) over the course of 10 game trials. Figure 3 shows the sequence of events for each trial in the user study. Participants are asked to decide whether to rely or not rely on the automation assistant prior to every trial, as shown in Figure 4(a). Regardless of their reliance choice, prior to the first trial, each participant is randomly assigned to one of the three $\Theta$ sets. Then, for their first five trials, a single $\theta_1$ value is randomly selected within the given $\Theta$ set. In this way, each participant experiences a constant input from the autonomous assistant for five repeated trials. Note that the participant is not informed of the specific $\theta$ value that is being applied to their input; they only know that the automation assistance is available and that they can turn it on or off. Moreover, for

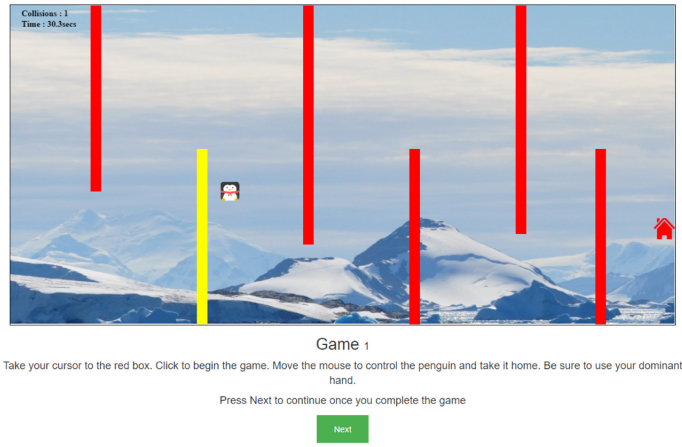



Fig. 2. A screenshot of the web-deployed game platform in which the participant must guide a penguin across the computer screen to its home while avoiding obstacles placed in its path.

Table 3. Game Parameters

| **Parameter** | $x_0$ | V | $\Delta t$ | $\theta_0$ | | $\phi(y)$ |
|---|---|---|---|---|---|---|
| | | | | | 0.75, | $y < \frac{1}{3}y_{max}$ |
| **Value** | [0, 200] | 75 pixel/sec | 0.02 sec | 1 | 1.25, | $\frac{1}{3}y_{max} \leq y < \frac{2}{3}y_{max}$ |
| | | | | | 1.75, | $y \geq \frac{2}{3}y_{max}$ |

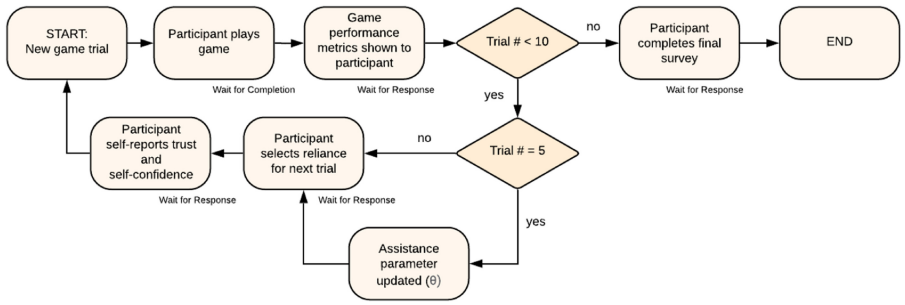
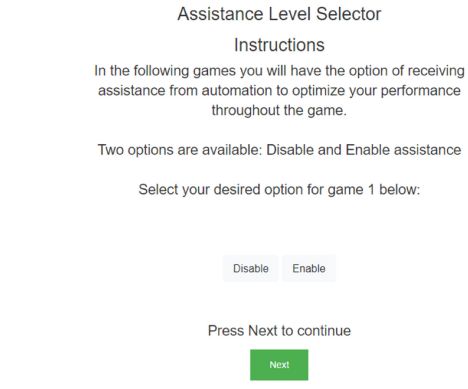



Fig. 3. The sequence of events in the experiment. The participant completes a practice trial prior to completing 10 trials of the game.

any game trial for which they choose not to rely on the automation assistant, $\theta_k = 1$. Similarly to [43], after each trial, participants are prompted to rate their trust (in the automation assistant) and self-confidence as shown in Figure 4(b). Participants are provided with definitions of the cognitive states prior to rating their trust and self-confidence on a numerical scale of 0–100. Trust is defined as assured reliance on the character, ability, strength, or truth of someone or something. Self-confidence is defined as confidence in oneself and in one's powers and abilities. While both trust and self-confidence self-report data are collected, only self-confidence self-report is used explicitly as an observation in the POMDP/R model as described in Section 3. Self-reported trust is utilized in validating the model's predictive capability.

At the sixth trial, a step change in the $\Theta$ set is introduced. The purpose of this step change is to further stimulate changes in the participant's trust or self-confidence. Note that to avoid




(a) Reliance Selection Page


(b) Survey Page

Fig. 4. Example screenshots of the survey 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 to 100.

introducing too large of a step change for some participants relative to others, no participant for whom $\theta_k \in \Theta_L \vee \Theta_H$ for trial $k = \{1, 2, 3, 4, 5\}$ experiences $\theta_k \in \Theta_L \vee \Theta_H$ for $k = \{6, 7, 8, 9, 10\}$. The choice of introducing the step change after five trials was based on analysis of data collected through pilot experiments. For the remaining five trials, a single $\theta_2$ value is then randomly selected within the new $\Theta$ set. Again, $\theta_k = 1$ for any trial $k$ during which the participant chooses not to rely on the automation assistant.

### 4.2 Implementation

A total of 367 individuals participated in, and completed, the study. These participants were recruited from the Amazon Mechanical Turk platform [1] 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 the online nature of the study, and given lack of participant supervision, it is assumed that some participants were not adequately

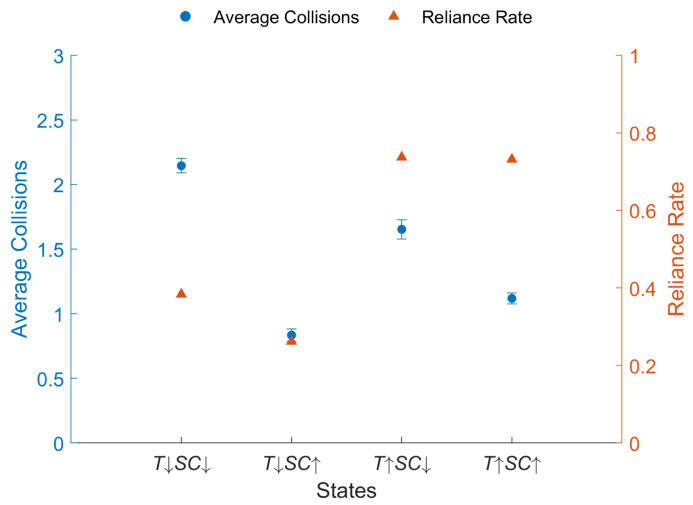



Fig. 5. Average collisions (left $y$-axis) and reliance rate (right $y$-axis) corresponding to the four combinations of trust and self-confidence, $T{\downarrow}SC{\downarrow}$, $T{\downarrow}SC{\uparrow}$, $T{\uparrow}SC{\downarrow}$, and $T{\uparrow}SC{\uparrow}$, as self-reported by participants. The error bars of the average collisions represent the standard error of the mean across participants.

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 25th percentile and with four or more collisions were removed. These conditions were chosen because they suggested that the participant dragged the penguin across the screen without attempting to avoid the obstacles. As a result, 27 participants were removed from the dataset. The resulting dataset consists of 340 participants from the United States (145 females, 190 males, 5 preferred not to disclose or did not identify within either gender), ranging in age from 18 to 77 (mean 39.0 and standard deviation 11.9, two participants did not disclose age).

### 4.3 Behavioral and Self-reported Data

Prior to training the POMDP/R model, the self-reported data is analyzed to identify behavioral trends. First, each participant's trust and self-confidence are identified as high or low by comparing the participant's self-reported value to the 50th percentile from all data. In Figure 5 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. On the right $y$-axis, 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 combination and the number of collisions and chosen reliance level of each participant associated with their reporting of each state. From Figure 5, it can be seen that the state combinations $T{\downarrow}SC{\downarrow}$ and $T{\uparrow}SC{\downarrow}$ correspond to poorer performance—i.e., greater average collisions. The established relationship between trust and reliance captured in previously published trust models is further underscored in Figure 5; when trust is high, the reliance rate is high, and vice versa. However, as expected, 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 both trust and self-confidence are high, $T{\uparrow}SC{\uparrow}$, it would have been expected that users would *not* rely on the assistant as often. However, participants who reported being in

Table 4. Estimator, P-values and Significance of Self-confidence Linear Regression Analysis

| | Estimate | p-value | Significance |
|---|---|---|---|
| Intercept | 65.8800 | 6.5870e − 233 | *** |
| Trial | 0.5140 | 1.095e − 04 | *** |
| Trust | 0.2377 | 3.6684e − 63 | *** |
| 1 Collision | −7.5242 | 6.2399e − 15 | *** |
| 2 Collisions | −15.1810 | 6.3625e − 37 | *** |
| 3 Collisions | −18.1590 | 5.8977e − 39 | *** |
| 4 Collisions | −23.4750 | 2.7911e − 41 | *** |
| 5 Collisions | −24.7240 | 4.5472e − 36 | *** |
| 6 Collisions | −21.3430 | 6.1672e − 18 | *** |
| Time | −0.2015 | 0.0171 | * |
| Automation Enabled | −4.8704 | 6.5045e − 06 | *** |
| $R^2$ | | 0.213 | |
| Adjusted $R^2$ | | 0.211 | |

*Note:* $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

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 Section 5.

### 4.4 Linear Regression Analysis

In order to further investigate the relationship between performance metrics and cognitive states, multi-variable linear regression analyses were applied to the data using the self-reported numerical self-confidence and trust data as regressors.

*Performance Metrics.* In Table 4, the estimated values show that as collisions and game time decrease, self-confidence increases. While all performance factors are significant for self-confidence, the categorical collision factors are much more significant to self-confidence than game time. This may be because as users progress through the trials and try to improve, avoiding obstacles is their priority. The intercept shown in Table 4 indicates that when automation is disabled and users have not collided with any obstacles, the baseline numerical self-confidence is 65.8800. In the trust regression analysis from Table 5, the estimates show that trust decreases when users collide with four to six obstacles and increases when users collide with one to three obstacles. Additionally, as game time increases, trust increases. Collisions are not found to be as significant to trust, whereas game time is; this may be because avoiding more obstacles typically implied that more time was spent navigating the penguin avatar across the screen. Additionally, the intercept in Table 5 suggests that the user avoiding all obstacles and having automation disabled is very significant to trust. Overall, these results suggest that self-confidence and trust have a positive relationship with absolute performance metrics as well as improving performance metrics.

*Cognitive States.* In both analyses, the corresponding cognitive state is also very significant. In other words, self-confidence is a significant factor of trust, and vice versa. Both numerical self-confidence and trust take on values of 0 to 100. Therefore, from the resulting regression estimate, a numerical trust rating of 100 translates to 23.77 points of self-confidence, and a numerical self-confidence rating of 100 translates to 33.53 points of trust. This is interesting because not only does this quantitatively suggest that self-confidence and trust affect each other, but also the relationship between trust and self-confidence is proportional. If trust and self-confidence are proportional to

Table 5. Estimator, P-values, and Significance of Trust Linear Regression Analysis

| | Estimate | p-value | Significance |
|---|---|---|---|
| Intercept | 10.1720 | 1.1573e − 04 | *** |
| Trial | −0.4378 | 0.0056 | ** |
| Self-Confidence | 0.3353 | 3.6684e − 63 | *** |
| 1 Collision | 1.0460 | 0.3634 | |
| 2 Collisions | 1.6460 | 0.2521 | |
| 3 Collisions | 0.2924 | 0.85124 | |
| 4 Collisions | −1.6333 | 0.4366 | |
| 5 Collisions | −1.4227 | 0.5481 | |
| 6 Collisions | −3.4316 | 0.2453 | |
| Time | 0.3073 | 0.0022 | * |
| Automation Enabled | 28.7120 | 7.2951e − 198 | *** |
| $R^2$ | | 0.310 | |
| Adjusted $R^2$ | | 0.308 | |

*Note:* $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

one another, the "confidence vs. trust" hypothesis may not be sufficiently able to predict reliance behavior when both cognitive states are high or low, thus further supporting the need for a model that does capture the nuances between trust and self-confidence. This proposed model is discussed in the next section.

## 5 MODEL TRAINING AND VALIDATION

The adaptation of the model to the specific HAI context considered in this article is first discussed in Section 5.1. This is followed by a description of the methods used for model training (Section 5.2) and model validation (Section 5.3).

### 5.1 Model Definition

Recall the T-SC cognitive state model defined in Table 2. 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 uncontrollable performance action set $\mathcal{A}_{uc,P}$ is further divided into tuples containing the number of *Collisions* $a_C$ and *Game Time* $a_G$, as shown in Equation (7). Additionally, the automation input $a_A$ is the assistance value $\theta$, discretized into the sets $\Theta_L, \Theta_M$, and $\Theta_H$ as described in Section 4 and referenced in Equation (8). Recall that $a_A$ is a controllable action in the context of the POMDP/R.

$$\mathcal{A}_{uc} = \{a_C, a_G\}$$

$$a_C \in C = \{\text{Collision Decrease } C^-,\ \text{Collision No Change } C^0,\ \text{Collision Increase } C^+\} \tag{7}$$

$$a_G \in G = \{\text{Game Time Decrease } G^-,\ \text{Game Time Increase } G^+\}$$

$$a_A \in \mathcal{A}_c = \{\Theta_L, \Theta_M, \Theta_H\} \tag{8}$$

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

$$\mathcal{T}(s'|s,a) = \mathcal{T}(s'_T|s_T, s_{SC}, a)\mathcal{T}(s'_{SC}|s_T, s_{SC}, a). \tag{9}$$

The emission probability function for reliance $\mathcal{E}_R : \mathcal{S} \times R \rightarrow [0,1]$ is represented by a $4 \times 2$ 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 $2 \times 2$ 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

$$\mathcal{E}(o|s) = \mathcal{E}(o_R|s_T, s_{SC})\mathcal{E}(o_{srSC}|s_{SC}). \tag{10}$$

Finally, the initial state probabilities for trust $\pi_T : T \rightarrow [0,1]$ and self-confidence $\pi_{SC} : SC \rightarrow [0,1]$ are both given by $1 \times 2$ matrices that represent the probability of the initial trust state $s_T$ and self-confidence state $s_{SC}$, respectively. As shown in Figure 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 152 effective parameters. There are 18 combinations of actions, consisting of the three collision performance distinctions, two game time performance distinctions, and three automation input value distinctions. There are four combinations of states, consisting of combinations of low and high levels of trust and self-confidence. Finally, there are four combinations of observations, consisting of the two levels of self-reported self-confidence as well as the two levels of reliance.

It should be noted that a limitation of the model is that the action space does not consider absolute performance. Ideally, the performance actions would be combinations of both change in performance and absolute performance. However, this would significantly increase the number of parameters in the model and, in turn, make model training computationally expensive. An analysis of models trained with performance defined either in absolute terms or as a delta between trials showed that the POMDP/R based upon change in performance actions leads to better predictability of the cognitive states and reliance behavior. Therefore, only change in performance is considered for the model presented here.

### 5.2 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 [5]. 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 [59]. Therefore, the genetic algorithm in MATLAB's Optimization Toolbox [2] is used to optimize the parameters of the model to maximize the likelihood of the sequences given the model parameters. The forward algorithm is used to calculate the likelihood of the sequences [60] in which the algorithm computes, recursively over time, the joint probability of a state $s_k$ at 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, $P(o_{1:N}, a_{1:N})$. The model was trained several times using randomized initialization. The resulting probabilities within each final trained model were identical up to at least four significant figures with a final log-likelihood of −3,446.4. Further model validation is included in Section 5.3.

Prior to training the model, the order of the action combinations and observation combinations is established. More specifically, the action combinations are ordered so that each of the transition probability matrices associated with these combinations can be distinguished prior to training. Similarly, the observation combinations are ordered for each of the emission probability matrices. In turn, this enables the state combination labels to be assigned a posteriori to the transition and emission probabilities, which ultimately enables interpretability and analysis of the trained probabilities. The assignment is based on the well-established trust-reliance relationship [24, 44] and context-specific knowledge, such as the expected likelihood of the human's self-reported self-confidence matching the model's prediction of self-confidence.

The state combination order of the resulting transition, emission, and initial probability matrices are 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 [42, 45, 51, 52]. After identifying the corresponding state combination of each row in the emission probability matrix, all rows and columns associated with states in the initial, transition, and emission probability matrices are re-ordered to match the prescribed state combination order.

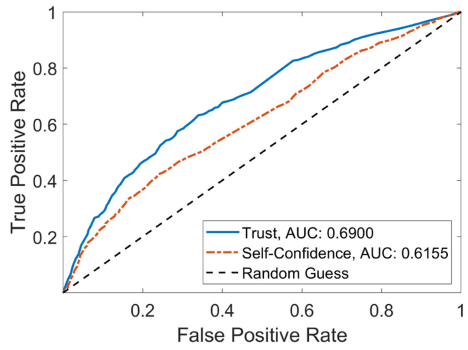
### 5.3 Validation

To test the predictive capability of the model and check for over-fitting, two validation methods are used. First, a $5 \times 2$-fold cross-validation is applied to the data in which the data is divided randomly into two equal sets, or folds. The model is trained with one fold and validated using the other. The entire process is then repeated for five iterations to increase the robustness of the validation log-likelihood values to variations in the training and testing datasets. The average log-likelihood of the trained models from 10-fold cross-validation is $-1{,}770.9 \pm 18.5$. In other words, the average log-likelihood of the $5 \times 2$-fold cross-validation varied by 1.1%, suggesting that the model is not overfitting the data.

Next, **receiver operating characteristic (ROC)** curves are utilized to illustrate the performance of the model in predicting the cognitive states and reliance decision of each participant. The cognitive state ROC curves (Figure 6(b)) are generated by comparing the self-reported cognitive states to the predicted belief state, as calculated using Equation (5), 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 shown in Figure 6(a), 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 (Figure 6(d)) 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 Figure 6(c). The model can predict both cognitive state levels and reliance choice better than a random guess as shown in Figures 6(b) and 6(d). 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

| | | 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) |

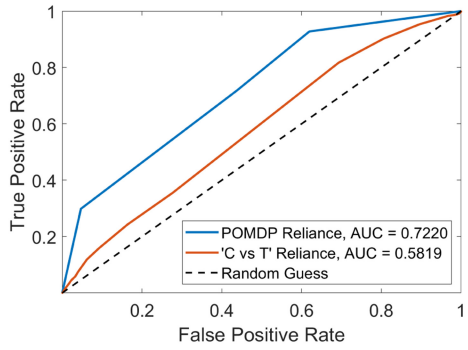
(a) Cognitive state confusion matrix




(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



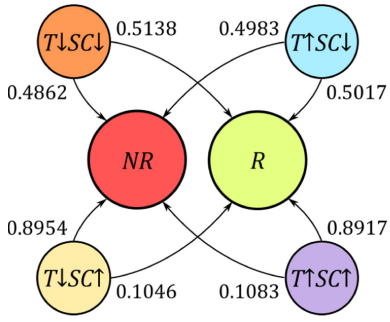

(d) ROC curve for reliance

Fig. 6. 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. The predicted reliance ROC curve using the "confidence vs. trust" hypothesis is also plotted in (d) and achieves an AUC of 0.58.

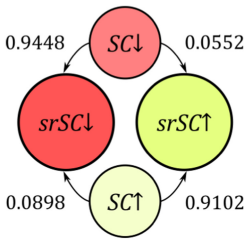
model classification performance. The trained model achieves a trust AUC of 0.69, self-confidence AUC of 0.62, and reliance AUC of 0.72.

### 5.4 Comparison Against "Confidence vs. Trust" Hypothesis

As discussed in Section 2, existing models of the relationship between human trust in automation, human self-confidence, and reliance on automation are based upon the "confidence vs. trust" hypothesis. Therefore, we compare the proposed model against that hypothesis. Using the self-reported trust and self-confidence values, an ROC curve using the "confidence vs. trust" hypothesis to predict participants' reliance behavior is generated and plotted in Figure 6(d). The true-positive rate is plotted against the false-positive rate using thresholds ranging from the minimum difference to the maximum difference between participants' self-reported trust and self-confidence. The ROC curve for predicted reliance using the "confidence vs. trust" hypothesis results in an AUC of 0.58 compared to that of the proposed model, which has an AUC of 0.72. From these results, we can conclude that the predictive capability of the proposed model, with respect to the user's reliance decision, is greater. From this metric alone, however, it is not possible to discern what aspect of the proposed model is responsible for this improvement in reliance prediction. Hence, differences between the model will be discussed more in Section 6.1.2.




(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.

Fig. 7. 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.

## 6 RESULTS AND DISCUSSION

In Section 6.1, 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 6.2) and a review of limitations (Section 6.3). Note that for ease of readability, the details of model parameters are provided in Appendix A. Moreover, note that to ensure that our model converged to a solution, 10 iterations of the POMDP/R were trained and the standard error of each parameter was found. It was found that the uncertainties of the initial, transition, and emission probabilities were considerably small compared to the parameter values themselves and that for several of the parameters, the standard error was found to be lower than the smallest value considered in MATLAB.

### 6.1 Results and Analysis

*6.1.1 Initial State Probabilities.* The initial state probabilities are provided in Table 7 (see Appendix A.1). From these probabilities it can be 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 toward automation, in which they trust automation prior to having any experience with it [24].

*6.1.2 Emission Probabilities.* Next the identified emission probabilities, visually depicted in Figures 7(a) and 7(b), are analyzed. Figure 7(a) shows the probability of reliance given the trust and self-confidence states, and Figure 7(b) shows the probability of self-reported self-confidence given the self-confidence state. The first observation from Figure 7(a) is that when the participant's 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 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. When they are in the $T\uparrow SC\uparrow$ state, they are highly likely (89.17%) to rely on it. 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 (51.38%) the automation assistance is nearly equally distributed. The same is true when participants are in the $T\uparrow SC\downarrow$ state. This suggests that self-confidence may be a more significant factor in reliance decisions when the user is in a state of low self-confidence rather than high self-confidence.

Table 6. Transition Probabilities for $a_A \in \Theta_L$, Decreasing Collisions, and Decreasing Time

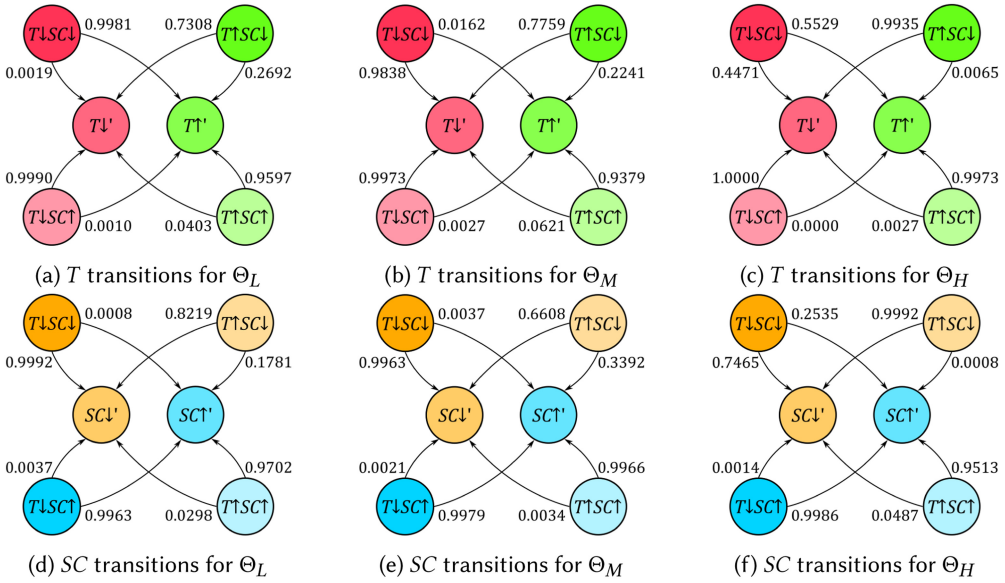
| | Trust | | Self-confidence | |
|---|---|---|---|---|
| | **T↓'** | **T↑'** | **SC↓'** | **SC↑'** |
| **T↓SC↓** | 0.0019 | **0.9981** | **0.9992** | 0.0008 |
| **T↓SC↑** | 0.9990 | 0.0010 | 0.0037 | **0.9963** |
| **T↑SC↓** | 0.7308 | 0.2692 | **0.8219** | 0.1781 |
| **T↑SC↑** | 0.0403 | 0.9597 | 0.0298 | **0.9702** |

It is also helpful to compare these probabilities directly to the reliance behavior predicted by models that build upon the "confidence vs. trust" hypothesis. The computational models discussed in Section 2 predict reliance based on a difference between the trust and self-confidence states. For example, using the hypothesis, it would be assumed that the $T{\uparrow}SC{\downarrow}$ state results in the participant relying and the $T{\downarrow}SC{\downarrow}$ results in them not relying on the automation. However, the emission probabilities shown in Figure 7(a) contradict this; instead, the likelihood of relying on or not relying on the automation, when self-confidence is low, is nearly 50%. It is worth noting that the proposed model is probabilistic, whereas existing ones are deterministic. Given the stochastic nature of human behavior, it is possible that the proposed model is able to better predict reliance behavior by inherently allowing for stochasticity in the prediction. In particular, it appears that when the human is in a state of low self-confidence, their behavior may be more stochastic than when they are in a state of high self-confidence. Recall the validation results shown earlier in Section 5.4 (see Figure 6(d)) in which the proposed model was a better predictor of reliance than a model based upon the "confidence vs. trust" hypothesis.

*6.1.3 Transition Probabilities.* Given that the POMDP/R 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 is 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 Figures 8(a) and 8(d), 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.* For easier interpretation, the referenced probabilities are in bold in Table 6.

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 Figures 8(a) to 8(f). They are likely to 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 toward themselves than the automation. Upon closer analysis, when $a_A \in \Theta_L \vee \Theta_M$, participants had a 26.91% and 22.41% chance, respectively, of remaining in a state of high trust, and a 17.81% and




(a) $T$ transitions for $\Theta_L$ (b) $T$ transitions for $\Theta_M$ (c) $T$ transitions for $\Theta_H$

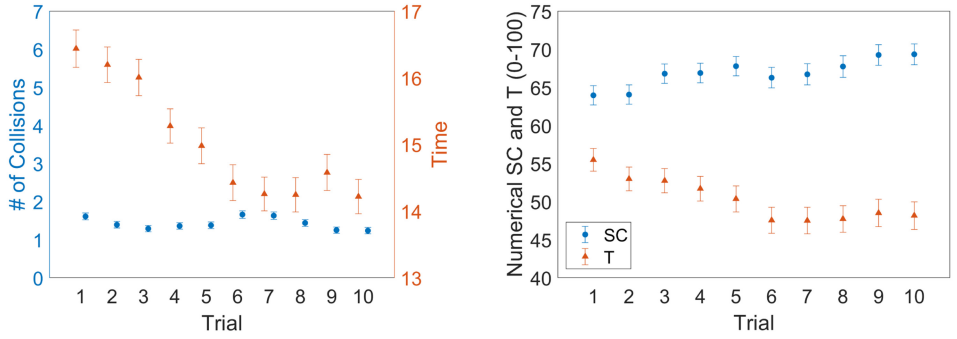
(d) $SC$ transitions for $\Theta_L$ (e) $SC$ transitions for $\Theta_M$ (f) $SC$ transitions for $\Theta_H$

Fig. 8. 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 $a_A \in \Theta_H$.

33.92% chance, respectively, of transitioning to a state of high self-confidence. 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 Figures 8(c) and 8(f), 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%) is 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 levels of automation assistance, 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 8 (see Appendix A.2), and when the participant is in the $T{\downarrow}SC{\downarrow}$ state, their likelihood of transitioning to a state of high trust (45.72%) or low 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 9. When $a_A \in \Theta_H$, as shown in Table 10, and the participant is in the $T{\downarrow}SC{\downarrow}$ state, they are highly likely (99.86%) to remain in a state of low




(a) Average number of collisions and average time taken per trial for all 10 trials

(b) Average numerical trust and self-confidence for all 10 trials

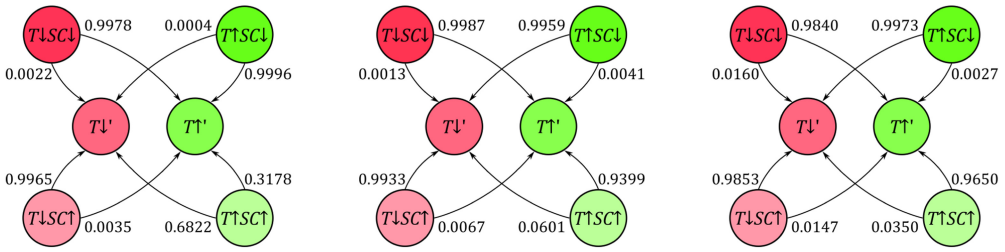
Fig. 9. Performance and self-reported trust and self-confidence over time.

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.

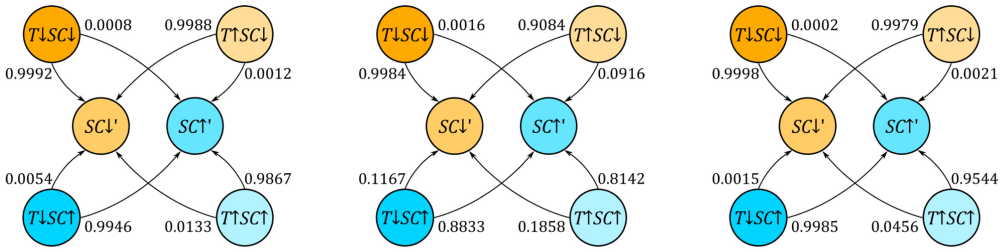
In comparing these results to the overall improvement case, participants in a state of low self-confidence are still unlikely to gain confidence and transition to $SC{\uparrow}$, but they are now not as likely to attribute any improvement to the automation. This underscores the consequences, from the 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.* An analysis of absolute collision and time performance data (see Figure 9(a)) shows that as the game progressed, on average, participants' performance improved and participants' self-confidence increased (see Figure 9(b)). In turn, these observations suggest that in addition to trust calibration, correct calibration of self-confidence is important for improved HAI, as discussed further in Section 6.2.

*Overall Performance Deterioration.* Next, cases in which participants' performance deteriorates between game trials are 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 $a_A \in \Theta_M \vee \Theta_H$ (Figures 10(b) and 10(c)), 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 participants' attribution of changes in performance, so does the user's experience with the autonomous assistant.

*Partial Performance Deterioration.* Next, the case in which the 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




(a) $T$ transitions for $\Theta_L$ (b) $T$ transitions for $\Theta_M$ (c) $T$ transitions for $\Theta_H$




(d) $SC$ transitions for $\Theta_L$ (e) $SC$ transitions for $\Theta_M$ (f) $SC$ transitions for $\Theta_H$

Fig. 10. 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 $a_A \in \Theta_H$.

in Tables 8–10, 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.

### 6.2 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 [71, 74]. 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 [44], in this case, participants' trust in the automation and their own 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 humans 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 [41, 44, 54, 73]. 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 [6, 77]. Finally, through the comparison of the AUCs from the reliance ROC curves, it was observed that the trained model outperforms the "confidence vs. trust" hypothesis. This supports the need for understanding the nuances between trust and self-confidence for the prediction of human reliance on automation.

### 6.3 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 article are specific to the HAI scenario under consideration. However, given the generalized definition of the POMDP/R 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 number 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 model adaptations for scenarios in which it does not, is another direction of future work. Future work may extend the given model to use a continuous state space to more accurately characterize trust and self-confidence dynamics. This would allow for incremental changes in these cognitive states to be accounted for [9]. Depending on the context, actions and observations may also be extended to the continuous space.

## 7 CONCLUSION

The contribution of this article 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 without a reward function 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 with coupled transition probability functions. By doing so, the model is able to capture the nuanced effects of various combinations of the states on the participant's reliance on autonomous assistance. A study 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 varies 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.

The takeaways of this work are as follows. First, attribution theory is critical when humans are interacting with automation. Second, the calibration of both trust and self-confidence is important to avoid misattributions of skills in HAI for learning contexts. Lastly, by accounting for the coupling between trust and self-confidence, the proposed model outperforms the "confidence vs. trust" hypothesis with respect to the prediction of human reliance on automation. This validates the need to understand the relationship between trust and self-confidence when humans decide to rely on or not rely on automation. Future work includes validation of the model for other HAI scenarios, investigation of individual differences that may lead to distinct trust or self-confidence dynamics, and model-based control algorithm design aimed at, for example, optimally allocating control authority to the human and automation based on calibration of the human's trust and self-confidence.

## APPENDIX

## A TRAINED MODEL RESULTS

We present the POMDP model of human trust–self-confidence behavior discussed in Section 5.

### A.1 Initial State Probabilities

The initial state probabilities for trust $\pi_T : 1 \times T \rightarrow [0, 1]$ and self-confidence $\pi_{SC} : 1 \times SC \rightarrow [0, 1]$ are both represented by $1 \times 2$ 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 7.

Table 7. 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

$$\mathcal{T}(s'|s, a) = \mathcal{T}(s'_T|s_T, s_{SC}, a)\mathcal{T}(s'_{SC}|s_T, s_{SC}, a). \tag{11}$$

The transition probabilities are provided in Tables 8–10. 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 8. 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 |

Table 9. 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 |

Table 10. 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 11. Emission Probabilities of the Reliance Observation $o_R$ and Self-reported Self-confidence Observation $o_{srSC}$

| | 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 | | | |

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.

### A.3 Emission Probabilities

The emission probability function for reliance $\mathcal{E}_R : \mathcal{S} \times R \rightarrow [0, 1]$ is represented by a $4 \times 2$ 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 $2 \times 2$ 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 reliance and self-reported self-confidence emission probabilities, given by

$$\mathcal{E}(o|s) = \mathcal{E}(o_R|s_T, s_{SC})\mathcal{E}(o_{srSC}|s_{SC}). \tag{12}$$

The emission probabilities are provided in Table 11.

## ACKNOWLEDGMENTS

We thank Sooyung Byeon (Purdue University) for initial development of the game platform that was adapted for the human subject experiment.

## REFERENCES

[1] 2018. Amazon Mechanical Turk. https://www.mturk.com/.
[2] 2021. MATLAB Optimization Toolbox. https://www.mathworks.com/products/optimization.html#resources.
[3] Kumar Akash. 2020. *Reimagining Human-machine Interactions through Trust-based Feedback*. Thesis. Purdue University Graduate School.
[4] Kumar Akash, Wan-Lin Hu, Neera Jain, and Tahira Reid. 2018. A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems* 8, 4 (Nov. 2018), 1–20. DOI: https://doi.org/10.1145/3132743 arXiv:1803.09861 [cs].
[5] Kumar Akash, Griffon McMahon, Tahira Reid, and Neera Jain. 2020. Human trust-based feedback control: Dynamically varying automation transparency to optimize human-machine interactions. *IEEE Control Systems Magazine* 40, 6 (Dec. 2020), 98–116. DOI: https://doi.org/10.1109/MCS.2020.3019151
[6] Kumar Akash, Tahira Reid, and Neera Jain. 2019. Improving human-machine collaboration through transparency-based feedback – Part II: Control design and synthesis. *IFAC-PapersOnLine* 51, 34 (Jan. 2019), 322–328. DOI: https://doi.org/10.1016/j.ifacol.2019.01.026
[7] Kumar Akash, Wan-Lin Hu, Tahira Reid, and Neera Jain. 2017. Dynamic modeling of trust in human-machine interactions. In *2017 American Control Conference (ACC'17)*. IEEE, Seattle, WA, 1542–1548. DOI: https://doi.org/10.23919/ACC.2017.7963172
[8] Susan M. Astley. 2005. Evaluation of computer-aided detection (CAD) prompting techniques for mammography. *British Journal of Radiology* 78, Suppl_1 (Jan. 2005), S20–S25. DOI: https://doi.org/10.1259/bjr/37221979
[9] Hebert Azevedo-Sa, Suresh Kumaar Jayaraman, Connor T. Esterwood, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. 2020. Real-time estimation of drivers' trust in automated driving systems. *International Journal of Social Robotics* 13 (Sept. 2020), 1911–1927. DOI: https://doi.org/10.1007/s12369-020-00694-1
[10] Victoria A. Banks, Neville A. Stanton, and Catherine Harvey. 2014. Sub-systems on the road to vehicle automation: Hands and feet free but not 'mind' free driving. *Safety Science* 62 (Feb. 2014), 505–514. DOI: https://doi.org/10.1016/j.ssci.2013.10.014
[11] Woodrow Barfield and Thomas A. Dingus. 2014. *Human Factors in Intelligent Transportation Systems*. (1st ed.) Taylor & Francis, New York.
[12] Jayson G. Boubin, Christina F. Rusnock, and Jason M. Bindewald. 2017. Quantifying compliance and reliance trust behaviors to influence trust in human-automation teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 61, 1 (Sept. 2017), 750–754. DOI: https://doi.org/10.1177/1541931213601672
[13] Erik Brockbank, Haoliang Wang, Justin Yang, Suvir Mirchandani, Erdem Bıyık, Dorsa Sadigh, and Judith E. Fan. 2022. How do people incorporate advice from artificial agents when making physical judgments? (May 2022). DOI: https://doi.org/10.48550/arXiv.2205.11613
[14] Jerome R. Busemeyer and James T. Townsend. 1993. Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review* 100 (1993), 432–459. DOI: https://doi.org/10.1037/0033-295X.100.3.432
[15] Eric T. Chancey, James P. Bliss, Yusuke Yamani, and Holly A. H. Handley. 2017. Trust and the compliance–reliance paradigm: The effects of risk, error bias, and reliability on trust and dependence. *Human Factors* 59, 3 (May 2017), 333–345. DOI: https://doi.org/10.1177/0018720816682648
[16] Jessie Y. C. Chen and Peter I. Terrence. 2009. Effects of imperfect automation and individual differences on concurrent performance of military and robotics tasks in a simulated multitasking environment. *Ergonomics* 52, 8 (Aug. 2009), 907–920. DOI: https://doi.org/10.1080/00140130802680773
[17] Min Chen, Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. Planning with trust for human-robot collaboration. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction (HRI'18)*. Association for Computing Machinery, New York, NY, 307–315. DOI: https://doi.org/10.1145/3171221.3171264
[18] Ewart de Visser, Samuel Monfort, Ryan Mckendrick, Melissa Smith, Patrick McKnight, Frank Krueger, and Raja Parasuraman. 2016. Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied* 22 (Aug. 2016), 331–349. DOI: https://doi.org/10.1037/xap0000092
[19] Ewart de Visser, Richard Pak, and Tyler Shaw. 2018. From "automation" to "autonomy": The importance of trust repair in human-machine interaction. *Ergonomics* 61 (March 2018), 1–33. DOI: https://doi.org/10.1080/00140139.2018.1457725

[20] Ewart de Visser and Raja Parasuraman. 2011. Adaptive aiding of human-robot teaming: Effects of imperfect automation on performance, trust, and workload. *Journal of Cognitive Engineering and Decision Making* 5 (June 2011), 209–231. DOI: https://doi.org/10.1177/1555343411410160

[21] Ewart J. de Visser, Marieke M. M. Peeters, Malte F. Jung, Spencer Kohn, Tyler H. Shaw, Richard Pak, and Mark A. Neerincx. 2020. Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics* 12, 2 (May 2020), 459–478. DOI: https://doi.org/10.1007/s12369-019-00596-x

[22] Peter de Vries, Cees Midden, and Don Bouwhuis. 2003. The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies* 58, 6 (June 2003), 719–735. DOI: https://doi.org/10.1016/S1071-5819(03)00039-9

[23] J. D. Dodson. 1915. The relation of strength of stimulus to rapidity of habit-formation in the kitten. *Journal of Animal Behavior* 5 (1915), 330–336. DOI: https://doi.org/10.1037/h0073415

[24] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (June 2003), 697–718. DOI: https://doi.org/10.1016/S1071-5819(03)00038-7

[25] Mica R. Endsley. 2017. From here to autonomy: Lessons learned from human–automation research. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 59, 1 (Feb. 2017), 5–27. DOI: https://doi.org/10.1177/0018720816681350

[26] Karen M. Feigh, Michael C. Dorneich, and Caroline C. Hayes. 2012. Toward a characterization of adaptive systems: A framework for researchers and system designers. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 54, 6 (Dec. 2012), 1008–1024. DOI: https://doi.org/10.1177/0018720812443983

[27] Michael W. Floyd, Michael Drinkwater, and David W. Aha. 2015. Improving trust-guided behavior adaptation using operator feedback. In *Case-Based Reasoning Research and Development (Lecture Notes in Computer Science)*, Eyke Hüllermeier and Mirjam Minor (Eds.). Springer International Publishing, Cham, 134–148. DOI: https://doi.org/10.1007/978-3-319-24586-7_10

[28] US Air Force. 2010. *Report on Technology Horizons: A Vision for Air Force Science & Technology during 2010–203*. Technical Report. https://www.airuniversity.af.edu/AUPress/Book-Reviews/Display/Article/1194559/report-on-technology-horizons-a-vision-for-air-force-science-technology-during/.

[29] Ulrike Esther Franke. 2014. Drones, drone strikes, and US policy: The politics of unmanned aerial vehicles. *Parameters* 44, 1 (2014), 121–130.

[30] Amos Freedy, Ewart DeVisser, Gershon Weltman, and Nicole Coeyman. 2007. Measurement of trust in human-robot collaboration. In *2007 International Symposium on Collaborative Technologies and Systems*. 106–114. DOI: https://doi.org/10.1109/CTS.2007.4621745

[31] Ji Gao and John D. Lee. 2006. 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* 36, 5 (Sept. 2006), 943–959. DOI: https://doi.org/10.1109/TSMCA.2005.855783

[32] Michael. A. Goodrich and Mary L. Cummings. 2015. Human factors perspective on next generation unmanned aerial systems. In *Handbook of Unmanned Aerial Vehicles*, Kimon P. Valavanis and George J. Vachtsevanos (Eds.). Springer Netherlands, Dordrecht, 2405–2423. DOI: https://doi.org/10.1007/978-90-481-9707-1_23

[33] Peter A. Hancock, Richard J. Jagacinski, Raja Parasuraman, Christopher D. Wickens, Glenn F. Wilson, and David B. Kaber. 2013. Human-automation interaction research: Past, present, and future. *Ergonomics in Design* 21, 2 (April 2013), 9–14. DOI: https://doi.org/10.1177/1064804613477099

[34] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 3 (2015), 407–434. DOI: https://doi.org/10.1177/0018720814547570

[35] Mark Hoogendoorn, Syed Waqar Jaffry, Peter Paul Van Maanen, and Jan Treur. 2013. Modelling biased human trust dynamics. *Web Intelligence and Agent Systems* 11, 1 (Aug. 2013), 21–40.

[36] Wan-Lin Hu, Kumar Akash, Neera Jain, and Tahira Reid. 2016. Real-time sensing of trust in human-machine interactions. *IFAC-PapersOnLine* 49, 32 (Jan. 2016), 48–53. DOI: https://doi.org/10.1016/j.ifacol.2016.12.188

[37] Wan-Lin Hu, Kumar Akash, Tahira Reid, and Neera Jain. 2019. Computational modeling of the dynamics of human trust during human–machine interactions. *IEEE Transactions on Human-Machine Systems* 49, 6 (Dec. 2019), 485–497. DOI: https://doi.org/10.1109/THMS.2018.2874188

[38] Aya Hussein, Sondoss Elsawah, and Hussein Abbass. 2020. 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*. DOI: https://doi.org/10.24251/HICSS.2020.047

[39] Ion Juvina, Michael G. Collins, Othalia Larue, William G. Kennedy, Ewart De Visser, and Celso De Melo. 2019. Toward a unified theory of learned trust in interpersonal and human-machine interactions. *ACM Transactions on Interactive Intelligent Systems* 9, 4 (Oct. 2019), 24:1–24:33. DOI: https://doi.org/10.1145/3230735

[40] Ion Juvina, Christian Lebiere, and Cleotilde Gonzalez. 2015. Modeling trust dynamics in strategic interaction. *Journal of Applied Research in Memory and Cognition* 4, 3 (Sept. 2015), 197–211. DOI: https://doi.org/10.1016/j.jarmac.2014.09.004

[41] Christian Lebiere, Leslie M. Blaha, Corey K. Fallon, and Brett Jefferson. 2021. Adaptive cognitive mechanisms to maintain calibrated trust and reliance in automation. *Frontiers in Robotics and AI* 8 (2021). https://www.frontiersin.org/article/10.3389/frobt.2021.652776.

[42] John D. Lee and Neville Moray. 1992. Trust, control strategies and allocation of function in human-machine systems. *Ergonomics* 35, 10 (1992), 1243–1270. DOI: https://doi.org/10.1080/00140139208967392

[43] John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-computer Studies* 40, 1 (1994), 153–184. DOI: https://doi.org/10.1006/ijhc.1994.1007

[44] John D. Lee and Katrina A. See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46, 1 (2004), 50–80. DOI: https://doi.org/10.1518/hfes.46.1.50_30392

[45] Stephan Lewandowsky, Michael Mundy, and Gerard P. A. Tan. 2000. The dynamics of trust: Comparing humans to automation. *Journal of Experimental Psychology: Applied* 6, 2 (2000), 104–123. DOI: https://doi.org/10.1037/1076-898X.6.2.104

[46] Morten Lind. 1999. Plant modelling for human supervisory control. *Transactions of the Institute of Measurement and Control* 21, 4–5 (Oct. 1999), 171–180. DOI: https://doi.org/10.1177/014233129902100405

[47] Peter-Paul Maanen, Francien Wisse, Jurriaan Diggelen, and Robbert Jan Beun. 2011. 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*. IEEE, 287. DOI: https://doi.org/10.1109/WI-IAT.2011.117

[48] P. Madhavan and D. A. Wiegmann. 2007. Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science* 8, 4 (July 2007), 277–301. DOI: https://doi.org/10.1080/14639220500337708

[49] Dariusz Mikulski, Frank Lewis, Edward Gu, and Greg Hudas. 2012. Trust method for multi-agent consensus. In *Unmanned Systems Technology XIV*, Vol. 8387. SPIE, Baltimore, MD, 146–159. DOI: https://doi.org/10.1117/12.918927

[50] Bonnie M. Muir. 1987. Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies* 27, 5 (1987), 527–539. DOI: https://doi.org/10.1016/S0020-7373(87)80013-5

[51] Bonnie M. Muir. 1990. *Operators' Trust in and Use of Automatic Controllers in a Supervisory Process Control Task*. Ph.D. Thesis. University of Toronto, Toronto, ON, Canada.

[52] Bonnie M. Muir and Neville Moray. 1996. Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (March 1996), 429–460. DOI: https://doi.org/10.1080/00140139608964474

[53] Heather Neyedli, Justin Hollands, and Greg Jamieson. 2009. Human reliance on an automated combat ID system: Effects of display format. *Human Factors and Ergonomics Society Annual Meeting Proceedings* 53 (Oct. 2009), 212–216. DOI: https://doi.org/10.1518/107118109X1252444108002

[54] Kazuo Okamura and Seiji Yamada. 2020. Adaptive trust calibration for human-AI collaboration. *PLOS ONE* 15, 2 (Feb. 2020), e0229132. DOI: https://doi.org/10.1371/journal.pone.0229132

[55] Raja Parasuraman and Dietrich H. Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors* 52, 3 (June 2010), 381–410. DOI: https://doi.org/10.1177/0018720810376055

[56] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39, 2 (June 1997), 230–253. DOI: https://doi.org/10.1518/001872097778543886

[57] Jeffrey R. Peters, Vaibhav Srivastava, Grant S. Taylor, Amit Surana, MIguel P. Eckstein, and Francesco Bullo. 2015. Human supervisory control of robotic teams: Integrating cognitive modeling with engineering design. *IEEE Control Systems Magazine* 35, 6 (Dec. 2015), 57–80. DOI: https://doi.org/10.1109/MCS.2015.2471056

[58] Joelle Pineau and Geoffrey J. Gordon. 2007. POMDP planning for robust robot control. In *Robotics Research (Springer Tracts in Advanced Robotics)*, Sebastian Thrun, Rodney Brooks, and Hugh Durrant-Whyte (Eds.). Springer, Berlin, 69–82. DOI: https://doi.org/10.1007/978-3-540-48113-3_7

[59] Óscar Pérez, Massimo Piccardi, Jesús García, Miguel Ángel Patricio, and José Manuel Molina. 2007. Comparison between genetic algorithms and the Baum-Welch algorithm in learning HMMs for human activity classification. In *Applications of Evolutionary Computing*, Mario Giacobini (Ed.). Lecture Notes in Computer Science, Vol. 4448. Springer, Berlin, 399–406.

[60] Lawrence Rabiner and Biing-Hwang Juang. 1986. An introduction to hidden Markov models. *IEEE ASSP Magazine* 3, 1 (1986), 4–16. DOI: https://doi.org/10.1109/MASSP.1986.1165342

[61] Mat R. Abdul Rani Rani, Murray A. Sinclair, and Keith Case. 2000. Human mismatches and preferences for automation. *International Journal of Production Research* 38, 17 (Nov. 2000), 4033–4039. DOI: https://doi.org/10.1080/00207540050204894

[62] Victor Riley. 1996. Operator reliance on automation: Theory and data. In *Automation and Human Performance: Theory and Applications* (1st ed.), Raja Parasuraman and Mustapha Mouloua (Eds.). CRC Press, Mahwah, NJ, 19–35.

[63] Behzad Sadrfaridpour, Maziar Fooladi Mahani, Zhanrui Liao, and Yue Wang. 2018. Trust-based impedance control strategy for human-robot cooperative manipulation. In *ASME 2018 Dynamic Systems and Control Conference*. American Society of Mechanical Engineers, V001T04A015 (8 pages). DOI: https://doi.org/10.1115/DSCC2018-9170

[64] Hamed Saeidi and Yue Wang. 2015. Trust and self-confidence based autonomy allocation for robotic systems. In *2015 54th IEEE Conference on Decision and Control (CDC'15)*. IEEE, 6052–6057. DOI: https://doi.org/10.1109/CDC.2015.7403171

[65] Hamed Saeidi and Yue Wang. 2019. Incorporating trust and self-confidence analysis in the guidance and control of (semi)autonomous mobile robotic systems. *IEEE Robotics and Automation Letters* 4, 2 (April 2019), 239–246. DOI: https://doi.org/10.1109/LRA.2018.2886406

[66] Nathan E. Sanders and Chang S. Nam. 2021. Chapter 19 - Applied quantitative models of trust in human-robot interaction. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, 449–476. DOI: https://doi.org/10.1016/B978-0-12-819472-0.00019-8

[67] Thomas B. Sheridan. 1992. *Telerobotics, Automation, and Human Supervisory Control*. MIT Press, Cambridge, MA.

[68] Olivier Sigaud and Olivier Buffet. 2013. *Markov Decision Processes in Artificial Intelligence*. John Wiley & Sons, Hoboken, NJ. 2009048651

[69] Harold Soh, Yaqi Xie, Min Chen, and David Hsu. 2020. Multi-task trust transfer for human–robot interaction. *International Journal of Robotics Research* 39, 2–3 (March 2020), 233–249. DOI: https://doi.org/10.1177/0278364919866905

[70] Yudong Tao, Erik Coltey, Tianyi Wang, Miguel Alonso, Mei-Ling Shyu, Shu-Ching Chen, Hadi Alhaffar, Albert Elias, Biayna Bogosian, and Shahin Vassigh. 2020. Confidence estimation using machine learning in immersive learning environments. In *2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR'20)*. IEEE, 247–252. DOI: https://doi.org/10.1109/MIPR49039.2020.00058

[71] Kees van Dongen and Peter-Paul van Maanen. 2013. A framework for explaining reliance on decision aids. *International Journal of Human-Computer Studies* 71, 4 (April 2013), 410–424. DOI: https://doi.org/10.1016/j.ijhcs.2012.10.018

[72] Alan R. Wagner, Paul Robinette, and Ayanna Howard. 2018. Modeling the human-robot trust phenomenon: A conceptual framework based on risk. *ACM Transactions on Interactive Intelligent Systems* 8, 4 (Nov. 2018), 26:1–26:24. DOI: https://doi.org/10.1145/3152890

[73] Lu Wang, Greg A. Jamieson, and Justin G. Hollands. 2011. 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* 55, 1 (Sept. 2011), 375–379. DOI: https://doi.org/10.1177/1071181311551077

[74] Bernard Weiner. 1986. *An Attribution Theory of Motivation and Emotion*. Vol. 92. Springer-Verlag, New York, NY.

[75] Rebecca Wiczorek and Joachim Meyer. 2019. Effects of trust, self-confidence, and feedback on the use of decision automation. *Frontiers in Psychology* 10 (2019), 519. DOI: https://doi.org/10.3389/fpsyg.2019.00519

[76] Anqi Xu and Gregory Dudek. 2015. OPTIMo: Online probabilistic trust inference model for asymmetric human-robot collaborations. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*. ACM, New York, NY, 221–228. DOI: https://doi.org/10.1145/2696454.2696492

[77] X. Jessie Yang, Vaibhav V. Unhelkar, Kevin Li, and Julie A. Shah. 2017. Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction (HRI'17)*. Association for Computing Machinery, New York, NY, 408–416. DOI: https://doi.org/10.1145/2909824.3020230

Received 27 August 2021; revised 26 January 2023; accepted 6 April 2023