REIMAGINING HUMAN-MACHINE INTERACTIONS THROUGH TRUST-BASED FEEDBACK

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Kumar Akash

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2020

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Neera Jain, Chair

School of Mechanical Engineering

Dr. Inseok Hwang

School of Aeronautics and Astronautics

Dr. Robert W. Proctor

Department of Psychological Sciences

Dr. Tahira Reid Smith

School of Mechanical Engineering

Approved by:

Dr. Nicole Key

Head of the Graduate Program

Dedicated to my family

for their constant and unwavering support throughout this incredible journey

ACKNOWLEDGMENTS

I would like to extend my heartfelt gratitude and thanks to my advisor, Prof. Neera Jain, for her invaluable guidance and support throughout my journey at Purdue. Not only did she provide a conducive environment to perform and disseminate effective research, but she also ensured that I have all the necessary resources and knowledge required for this dissertation. I have learned a lot from her hardworking nature as well as encouraging and productivity-building approach towards her students. The freedom she has consistently extended to work on new ideas has been a great motivation for me. I am forever indebted to her for all the career advice and the numerous opportunities she has provided me with throughout the journey.

I am also incredibly thankful to Prof. Tahira Reid Smith for her great ideas, invaluable guidance, and advice throughout my research. Collaboration between the Jain Research Lab and the Reid Lab is great for fostering an incredible research environment, which I feel is second to none. I would also like to thank the members of my advisory committee—Prof. Inseok Hwang and Prof. Robert Proctor—for their valuable insights, constructive comments, and warm encouragement.

The work presented in this dissertation was not undertaken by me alone. I would like to acknowledge my talented collaborators, Dr. Wan-Lin Hu, Katelyn Polson, and Griffon McMahon for their contributions toward achieving the research goals. The members of the Jain Research Lab have contributed immensely to my personal and professional time at Purdue. I am grateful to Rian, Austin, Ana, Trevor, Aaron, Jianqui (Jack), Karan, Katie, and Matthew for the stimulating discussions, their unfailing support, feedback, and encouragement. Additionally, I would like to acknowledge the Herrick community for providing a warm and collaborative working environment.

A special thanks to Yeshaswi for her support, encouragement, and patience and for always helping me bring out my best. Finally, I am profoundly grateful to my

parents for all their love and continuous encouragement throughout my years of study and their support in all my pursuits. To my brother, Ankur Kumar, I simply wouldn't be where I am today without you. Thank you for always standing by me through thick and thin. I also thank Bhabhi and Riyansh for their love and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

ABSTRACT

Akash, Kumar Ph.D., Purdue University, August 2020. Reimagining Human-Machine Interactions through Trust-Based Feedback. Major Professor: Neera Jain, School of Mechanical Engineering.

Intelligent machines, and more broadly, intelligent systems, are becoming increasingly common in the everyday lives of humans. Nonetheless, despite significant advancements in automation, human supervision and intervention are still essential in almost all sectors, ranging from manufacturing and transportation to disaster-management and healthcare. These intelligent machines *interact and collaborate* with humans in a way that demands a greater level of trust between human and machine. While a lack of trust can lead to a human's disuse of automation, over-trust can result in a human trusting a faulty autonomous system which could have negative consequences for the human. Therefore, human trust should be *calibrated* to optimize these human-machine interactions. This calibration can be achieved by designing human-aware automation that can infer human behavior and respond accordingly in real-time.

In this dissertation, I present a probabilistic framework to model and calibrate a human's trust and workload dynamics during his/her interaction with an intelligent decision-aid system. More specifically, I develop multiple quantitative models of human trust, ranging from a classical state-space model to a classification model based on machine learning techniques. Both models are parameterized using data collected through human-subject experiments. Thereafter, I present a probabilistic dynamic model to capture the dynamics of human trust along with human workload. This model is used to synthesize optimal control policies aimed at improving context-specific performance objectives that vary automation transparency based on human state estimation. I also analyze the coupled interactions between human trust and

workload to strengthen the model framework. Finally, I validate the optimal control policies using closed-loop human subject experiments. The proposed framework provides a foundation toward widespread design and implementation of real-time adaptive automation based on human states for use in human-machine interactions.

1. INTRODUCTION

Automation has become prevalent in the everyday lives of humans. However, despite significant technological advancements, human supervision and intervention are still necessary in almost all sectors of automation, ranging from manufacturing and transportation to disaster-management and healthcare [1]. Therefore, we expect that the future will be built around *human-agent collectives* [2] that will require efficient and successful interaction and coordination between humans and machines. It is well established that to achieve this coordination, human trust in automation plays a central role [3–5]. For example, the benefits of automation are lost when humans override automation due to a fundamental lack of trust [3,5], and accidents may occur due to human mistrust in such systems [6]. Therefore, trust should be appropriately *calibrated* to avoid disuse or misuse of automation [4].

These negative effects can be overcome by designing autonomous systems that can adapt to a human's trust level. One way to adapt automation based on human trust is to augment the user interface with more information—either raw data or processed information—to help the human make an informed decision. This "amount of information" has been defined as automation *transparency* in the literature. Transparency has been defined as "the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process" [7]. With higher levels of transparency, humans have access to more information to aid their decisions, which has been shown to increase trust [8, 9]. Therefore, adaptive automation can be implemented by varying automation transparency based on human trust. However, because high transparency requires communicating more information, it can also increase the workload of the human [10]. In turn, high levels of workload can lead to fatigue and therefore reduce the human's performance. Therefore, we need to design an adaptive automation

that can vary automation transparency to accommodate changes in human trust and workload in real-time to achieve optimal or near-optimal performance. *This requires a dynamic model of human trust-workload behavior to design and implement control policies.*

1.1 Background

Human trust is a multidisciplinary concept. Each discipline characterizes a different type of relationship with the term "trust". While disciplines like social sciences often study the trust between individuals or organizations, other fields like computing and networking evaluate trust with respect to artificial intelligence and communication networks; in other words, the definition of trust varies drastically. Nonetheless, in most disciplines, trust captures the interaction between two entities, a trustor who relies on a trustee in a situation consisting of uncertainty and risk. Cho et al. summarized trust across disciplines as "the willingness of the trustor to take risk based on a subjective belief that a trustee will exhibit reliable behavior to maximize the trustor's interest under uncertainty of a given situation based on the cognitive assessment of past experience with the trustee [11]." With respect to automation, trust is defined as "the attitude that one agent will achieve another agent's goal in a situation where imperfect knowledge is given with uncertainty and vulnerability [4]." For an interaction between a human and an automated system, the human expects and therefore trusts the automated system to achieve a desired goal in an uncertain and risky environment.

In the context of autonomous systems, human trust can be classified into three categories: dispositional, situational, and learned [12]. Dispositional trust refers to the component of trust that is dependent on demographics such as gender and culture, whereas situational and learned trust depend on a given situation (e.g., task difficulty) and past experience (e.g., machine reliability), respectively. While all of these trust factors influence the way humans make decisions when interacting with

automation, situational and learned trust factors "can change within the course of a single interaction" [12]. Therefore, I am interested in using feedback control principles to design adaptive automation that is capable of calibrating the human's trust in the automation in real time so that it maintains a successful interaction; here, "success" is context specific.

1.2 Levels of Automation and Transparency

Automation can be differentiated by levels, with higher levels of automation (LOAs) representing increased machine autonomy. Parasuraman et al. proposed that most systems involve four stages of sequential tasks, with each successive stage dependent on successful completion of the previous one. They correspond to the four information-processing stages of humans: (1) information acquisition (sensory processing), (2) information analysis (perception), (3) decision and action selection (decision-making), and (4) action implementation (response selection) [13, 14]. This four-stage model of human information processing has its equivalence in system functions that can be automated, leading to four types of automation as shown in Figure 1.1. These can be conveniently called acquisition, analysis, decision, and action automation, respectively. Within this four-stage model, each type of automation can have different degrees or levels of automation (LOA), depending on the context.

Information sources are inherently uncertain due to factors including sensor imprecision and unpredictable events. Nevertheless, even imperfect acquisition and analysis automation at a reliability as low as 70%, can improve performance as compared to unaided human performance [15]. On the other hand, the utility of *decision* and *action* automation is more sensitive to adequately calibrated human trust in the automation than *acquisition* and *analysis* automation. This is because in the case of decision automation, the human is typically being asked to comply, or not comply, with the decision proposed by the automation. Similarly, in the case of action automation, the human is being asked to rely, or not rely, on the actions being taken

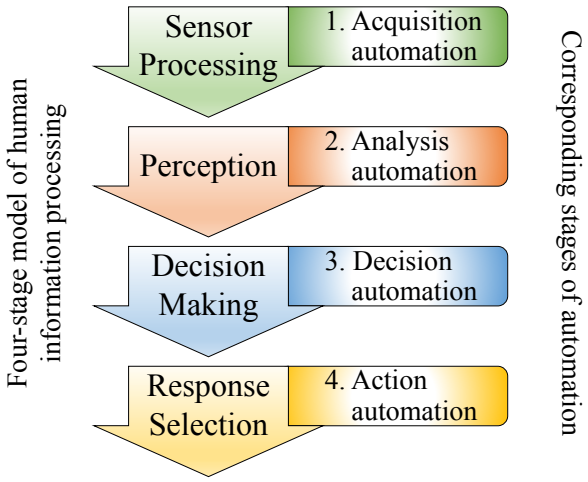


Figure 1.1. Simple four-stage model of human information processing and the corresponding types of automation (adapted from [13]). Each stage of human information processing has its equivalent in system functions that can be automated, thereby leading to four types of automation.

by the automation. In both cases, how strongly the human *trusts* the automation will affect their compliance with the automation (or the lack thereof). This can be dangerous in cases when the human's trust is not adequately calibrated to the reliability of the automation, such as when a human trusts an automation's faulty decision. It is possible to mitigate this issue by varying the LOA depending on situational demands during operational use; this has been classically defined as adaptive automation [16–18]. For decision-aid systems, adaptive automation can be realized by adjusting, or controlling, the amount of raw or processed information given to the human—in other words, controlling *transparency*. With higher levels of transparency, the operator has access to more information from lower LOAs, which has been shown to increase human trust [8, 9]. See [19] for a review on system transparency during human automation interactions.

1.3 Effects of Transparency on Trust and Workload

Several researchers have investigated the effect of transparency on trust, where systems with higher transparency have been shown to enhance humans' trust in systems [20–22]. Studies described in [23–25] showed that only the robot's reliability influenced trust; however, the studies also highlighted the limitation of the use of self-reported trust data. Notably, researchers have argued that high transparency can even reduce trust if the information provided is not interpretable or actionable to humans [26]. Furthermore, too much detail communicated through higher transparency can increase the time required to process the information [20] and distract the human from critical details [27]. Researchers have shown that cognitive difficulty, and thereby cognitive workload, increases with an increase in information [28]. Findings in [9] suggest that increased system transparency increases trust in the system but also causes workload to increase. Although a recent study based on the findings of [9] found that workload need not necessarily increase with greater transparency, several researchers agree that a trade-off between increased trust and increased workload exists when considering increased transparency [19].Typically, for decision-aid systems, this trade-off can be observed via the speed-accuracy tradeoff (SAT). Since trust calibration results in better decision making during human-machine interactions, the accuracy of human responses is better when human trust is calibrated. Moreover, increased workload due to an increase in transparency is often characterized by a longer response time required for processing more information [9, 29]; therefore, human response time also increases with an increase in transparency. SAT is a well-studied problem with established theoretical frameworks as well as neurobiological basis [30, 31]. It has been shown that, for simple stimuli, humans rapidly adjust their SAT to maximize rewards; however, humans deviate from the optimal SAT in complex situations [32]. An optimal choice of automation transparency based on human trust and workload estimates can aid in addressing the SAT while interact-

ing with decision-aid systems. Therefore, to optimally vary automation transparency, I model both human trust as well as human workload in a quantitative framework.

1.4 Dissertation Objectives

My main contribution in this dissertation is the design of a trust-based feedback control framework to improve context-specific performance objectives using automation transparency during human-machine collaborations. The overall framework is represented by Figure 1.2. Achieving this requires a predictive, reliable, and quantitative framework to estimate and calibrate human trust in automation. I demonstrate that human trust can be modeled and predicted based on a human's behavioral responses to the automation's decision-aids. I also develop a model to estimate human trust using their psychophysiological measurements. I subsequently present the use of the modeled human trust and workload dynamics to optimally vary automation transparency by closing the loop between human and machine. In this dissertation, I will restrict the work to decision automation only, but the framework can be extended to action automation as well.

The following two sub-objectives will be achieved in order to satisfy the dissertation objective:

Objective 1: ***Dynamic Modeling and Estimation of Human Trust and Workload.*** Develop control-oriented models to estimate and predict the dynamics of human trust and workload behavior during an interaction with an autonomous system.

Objective 2: ***Human State-based Feedback Control.*** Synthesize a closed-loop controller that calibrates human trust and reduces workload to improve a context-specific performance objective during human-machine interactions.

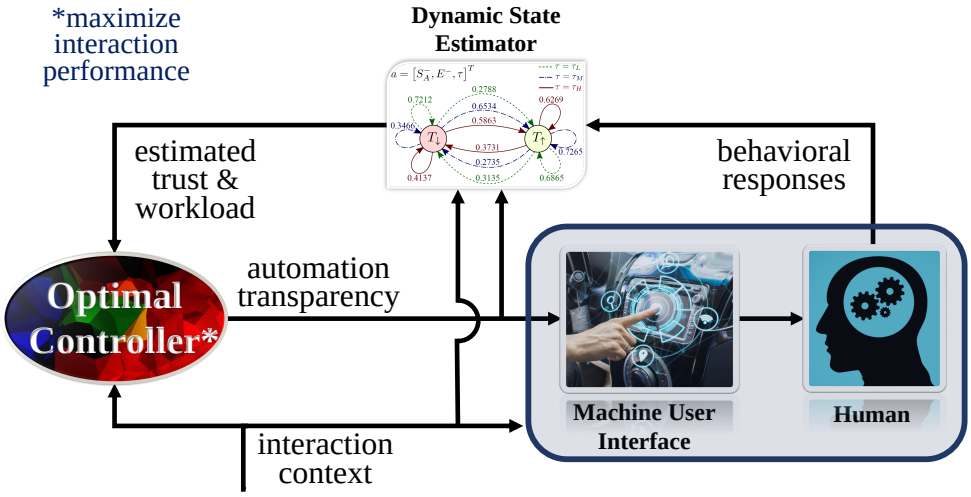


Figure 1.2. Block diagram depicting a trust and workload-based feedback control architecture for optimizing human-machine interactions. The human behavior model is used to estimate the non-observable human states of trust and workload using the machine outputs, the interaction context, and the observable human responses. An optimal control policy dynamically varies automation transparency based on the estimated human states to maximize a context specific performance objective.

1.4.1 Objective 1: Dynamic Modeling and Estimation of Human Trust

Researchers have studied trust behavior in human-machine interactions (HMI) [3, 33, 34] and analyzed the statistical significance of demographic factors (e.g., age, gender) on trust behaviors [35, 36]. However, while identifying factors that induce changes in trust is a critical step towards characterizing trust behavior, it is alone insufficient for characterizing a *quantitative* model of this behavior. Moreover, studies have shown that the trust level of humans varies with time due to changing experiences [33, 37] and, as such, any quantitative trust model should be dynamic.

In order to derive a quantitative dynamic model of human trust behavior suitable for HMI contexts, an appropriate experimental design, modeling approach, and model verification are necessary. To accomplish Objective 1, I first develop the modeling and experimental methods to capture *dynamic changes* in human trust with varying automation reliability, specifically in a Stage 3 automation context. I also systemat-

ically analyze the effects of demographic factors, consisting of national culture [38] and gender, as well as system error type.

Furthermore, in some uncertain and unstructured environments, it is not practical to retrieve human behavior continuously for use in a feedback control algorithm. Specifically, in uncertain environments, we may not always be able to associate human responses to certain behaviors. In other instances, human responses can simply be lost due to faults in communication. Similarly, in unstructured environments, it may not be possible to characterize specific cognitive behaviors of a human (e.g., trust) based upon observations of their physical behavior or actions alone. In such scenarios, an alternative is the use of psychophysiological signals to estimate trust level [39]. While these measurements have been correlated to human trust level [40, 41], they have not been studied in the context of real-time trust sensing.

There are few psychophysiological measurements that have been studied in the context of human trust. I focus here on electroencephalography (EEG) and galvanic skin response (GSR) which are both noninvasive and whose measurements can be collected and processed in real time. EEG is an electrophysiological measurement technique that captures the cortical activity of the brain [42]. Some researchers have studied trust via EEG measurements [40] but their methodology is infeasible for real-time trust level sensing. GSR is a classical psychophysiological signal that captures arousal based upon the conductivity of the surface of the skin. It is not under conscious control but is instead modulated by the sympathetic nervous system. Researchers have examined GSR in correlation with human trust level [43]. However, the use of GSR for *estimating* trust has not been explored and was noted as an area worth studying [39]. With respect to both GSR and EEG, a fundamental gap remains in determining a static model that not only estimates human trust level using these psychophysiological signals but that is also suitable for real-time implementation. In this dissertation, I develop a framework for a human trust sensor model using these psychophysiological measurements. The model is based upon data collected through a human subject study and the use of classification algorithms. I also present a prelimi-

nary sensor fusion technique to combine the psychophysiological measurements-based and behavioral data-based trust estimates.

In spite of the fact that 1) human trust is strongly dependent on automation reliability and 2) the modeled relationship between reliability and trust can be used to predict trust, autonomous systems cannot (and should not) vary their reliability to affect human trust. Instead, automation *transparency* can be varied to affect human trust. Moreover, although higher transparency can increase human trust [8, 9], it also can increase human workload [10, 28]. Therefore, to complete Objective 1, I model the dynamic effects of machine transparency on human trust and workload behavior so that it can be used for improving human-machine collaboration. Additionally, I develop models that explore multiple degrees of coupled interactions between human trust and workload and analyze their performance.

1.4.2 Objective 2: Human State-based Feedback Control

Although researchers have developed various models of human trust [44–47] and workload [48, 49], there does not exist a closed-loop framework for influencing human trust and workload to improve human-machine collaboration. Furthermore, published studies have shown that transparency affects both human trust [8, 9] and workload [10, 28] but has not been systematically used to calibrate trust-workload behavior. Therefore, a fundamental gap remains in using machine transparency to dynamically improve human-machine collaboration.

With the dynamic models of human trust and workload developed in Objective 1, the human state of trust and workload can be estimated as well as predicted. Using the developed models, I establish a systematic method for synthesizing a feedback controller to close the loop between human and machine. I design and synthesize control policies that vary machine transparency based on real-time estimation of human trust and workload to improve a context-specific performance metric. These control policies are validated using a reconnaissance mission study in which human

subjects are aided by a virtual robotic assistant. I show that this framework provides a tractable methodology for using human behavior as a real-time feedback signal to optimize human-machine interactions through dynamic modeling and control.

1.5 Outline

The rest of the document is organized as follows. In Chapter 2, I describe methodologies to model and estimate human trust using behavioral as well as psychophysiological measurements. In Chapter 3, I present a probabilistic framework to model human trust and workload dynamics. In addition, I present closed-loop policies that vary machine transparency based on real-time estimation of human trust and workload to improve a context-specific performance metric. I further extend this framework by considering the coupled interactions between human trust and workload in Chapter 4. Finally, conclusions are drawn from this work in Chapter 5 and potential future work is discussed.

2. DYNAMIC MODELING AND ESTIMATION OF HUMAN TRUST

Researchers have studied trust behavior in human-machine interactions (HMI) and human-computer interactions (HCI) using experimental methods and modeling techniques from social psychology [3,33,34]. Some studies focused on analyzing the statistical significance of demographic factors (e.g. age, gender) on trust behaviors [35,36]. However, while identifying factors that induce changes in trust is a critical step towards characterizing trust behavior, it is alone insufficient for characterizing a *quantitative* model of this behavior. Moreover, studies have shown that the trust level of humans varies with time due to changing experiences [33, 37] and, as such, any quantitative trust model should be dynamic. In Section 2.1, we will first develop a quantitative dynamic model of human trust that captures the effects of automation reliability and error type (miss or false alarm). We will also systematically analyze the effects of demographic factors, consisting of national culture [38] and gender, as well as system error type on human trust dynamics. Thereafter, in Section 2.2, we will develop a classification model to estimate human trust using psychophysiological measurements, specifically EEG and GSR. Finally, we will present an adaptive classification framework that will combine psychophysiological measurements and human behavioral dynamics to estimate human trust in real time in Section 2.3.

2.1 Modeling Effects of Automation Reliability

The contents of this section were previously published by Hu, Akash, Reid, and Jain in *IEEE Transactions on Human-Machine Systems* [47] and are reported here with minor modifications.

2.1.1 Introduction

In order to derive a quantitative dynamic model of human trust behavior suitable for HMI contexts, an appropriate experimental design, modeling approach, and model verification are necessary. There is no experimentally verified model for describing the dynamics of human trust level in HMI contexts that (1) incorporates demographic factors and time-varying experiences and (2) is built on experiments that elicit multiple transitions in trust level. Existing quantitative models either assume that human trust behavior is fully based on rationale [33] or are nonlinear [37, 50]. While the influence of accumulated effects of past interactions on the future trust level have been modeled in multi-agent system contexts, they have not been modeled for independent human-machine interactions [50,51]. Furthermore, existing models of human trust in autonomous systems have not taken into account human bias nor attitudes toward the system response bias; here the *system response bias* is defined in terms of signal detection theory, i.e., liberal (false-alarm-prone) and conservative (miss-prone) system bias.

Finally, human behavior is highly influenced by one's surroundings and past experiences [52, 53] with the automation's reliability, which in turn, are strongly influenced by demographic factors. With the spread of automation across the globe, it is necessary to model human behavior for different demographics. Several types of autonomous systems such as cars, smart thermostats, and tour-guide robots are designed to interact with unspecified users. In these contexts, a model that describes the trust dynamics of a population (general or grouped by demographics) instead of an individual would facilitate the design of such systems. Unfortunately, a generalized model that is suitable for capturing these variations in human trust behavior does not exist in the current literature.

In this section, we describe the modeling and experimental methods used to capture *dynamic changes* in human trust, specifically in a HMI context. We devise two sets of human subject experiments that elicit multiple dynamic transitions in human

trust behavior, and the data collected from these experiments is then used to estimate and validate the parameters of the proposed model. We establish a linear model motivated by literature on computational models for the dynamic variation of human trust [33, 54]; the linearity allows for easier control analysis and synthesis aimed at designing adaptive human-machine interfaces, thus enabling autonomous systems to respond to human trust variations in real-time. Furthermore, we systematically analyze the effects of demographic factors, consisting of national culture [38] and gender, as well as system error type.

This section is organized as follows. First, we provide background on trust modeling and significant factors that affect human trust. In Section 2.1.3, we describe experiment 1 along with the development of a generalized model of human trust dynamics; we further examine the effects of national culture and gender on these dynamics. In Section 2.1.4, we describe experiment 2 in which we incorporate the effects of misses and false alarms into the model. Afterward, we discuss the implications of the estimated parameter values of the trust model in the context of HMI.

2.1.2 Background

Comprehensive reviews of the influence of trust in HMI and HCI contexts are provided by Lee and See [4] and Hoff and Bashir [12]. Hoff and Bashir classified trust into three categories: dispositional, situational, and learned [12]. *Dispositional trust* is based on characteristics of the human. Factors that influence dispositional trust do not vary with time, but they still impact human decision-making during interactions with the autonomous system. Studies have shown differences in trust behavior between people of different cultures, age groups, and personality types [55–57]. *Situational trust* consists of factors that are external to the operator (e.g., task difficulty, potential risks) and those that are internal to the operator (e.g., self-confidence, domain knowledge) [12]. Finally, *learned trust* is based upon an operator's overall experience with an autonomous system and influences their initial mindset. During

a new interaction with an autonomous system, the human's experience affects their established trust level. In this section, we present a dynamic model that can capture variations in human trust level with respect to automation reliability for various demographic groups, thus, capturing both learned and dispositional trust characteristics.

In the remainder of this section, we first introduce studies that have modeled human trust in various contexts. Second, we review literature on determining the effect of automation reliability and system error types (i.e., misses or false alarms) on human trust. These factors influence learned trust. Last, we review studies on examining dispositional trust factors, specifically gender and national culture.

Studies on Human Trust Modeling

Researchers have developed models for predicting human trust based on past experiences, which strongly influence learned trust. Jonker and Treur suggested two types of functions to model trust dynamics: trust update functions and trust evolution functions [33]. Trust update functions use the current trust level and current experience to update the future trust level, while trust evolution functions map a sequence of trust related events (experiences) to a current trust level. In order to verify the proposed trust dynamics, Jonker et al. conducted follow-up human subject experiments which presented participants with a sequence of short stories for two scenarios: a photocopier and a travel agency [58]. Each scenario consisted of five positive and five negative stories, and participants reported their trust level after reading each story. The results suggested that trust dynamics are dependent on positive and negative experiences. However, limited by the number of trials (10 trials in each scenario), these studies only induced a single transition in trust level; therefore, the model did not capture the variations in trust dynamics involving multiple transitions.

Some studies have modeled human trust in the context of HMI. Lee and Moray studied changes in human trust level using a simulated semi-automatic juice plant

environment. It was observed that the human trust level was affected by the performance of the system, past trust levels, and faults [34]. The authors used an ARMAV (Auto Regressive Moving Average Vector) analysis to model the input-output relationship of the trust behavior. They later showed that humans use automation when their trust in the automation exceeds their self-confidence [59]. These early efforts demonstrated the effect of situational and learned trust on the interactions between humans and autonomous systems. However, due to a small sample size (i.e., four to five participants in each group) and a large standard deviation of the data, the accuracy of their model was limited.

Within the simulation context proposed by Moray and his colleagues, Lewandowsky et al. compared trust in automation with trust in human partners in equivalent situations [60]. Similar to the findings of Lee and Moray [59], Lewandowsky et al. identified that faults in auxiliary control actions have a strong effect on trust and self-confidence of the human operator, and the difference between trust and self-confidence is a strong predictor of the human operator's reliance on automation as well as his/her reliance on human colleagues.

Factors that are significant in predicting trust level may also be dependent on the application context. Sadrfaridpour et al. proposed a time-series model for the dynamics of human-robot trust in assembly lines based on the robot performance, human performance, and fault occurrences [61]. More specifically, the performances were quantified by the robot working speed and the human's state of muscle fatigue and recovery. How well the robot met the human's pace influenced the workload and trust perceived by the human. The researchers' experimental results also suggested that the current trust is mainly dependent on the previous trust if there is no dramatic change in performance.

More recently, elements that are not based on rationale have been incorporated into a human trust model. Li et al. used the structural equation modeling technique to identify the significance of human attitudes and subjective norm on "trusting intentions" [62]. Hoogendoorn et al. developed models with biased experience and/or

trust to account for this human behavior [37]. They validated their models using a geographical area classification task and showed that a model with a bias term is capable of estimating trust more accurately than models without an explicit bias. However, their model was nonlinear in trust and experience, rendering it more difficult for analysis than linear models.

The Effect of Misses and False Alarms in Automation

Automation reliability significantly influences human trust in autonomous systems and in turn influences human use of these systems [34, 63]. According to signal detection theory, automation errors can be classified as misses or false alarms; failing to detect the presence of a signal constitutes a miss, and incorrectly alerting humans to an absent stimulus constitutes a false alarm [64]. Existing literature shows that these two types of errors have different effects on human trust in automation. Specifically, these error types affect *reliance* and *compliance* to a different degree. Reliance is when humans, in the absence of any signal from the system, continue to trust the system and refrain from a response. On the other hand, compliance is exhibited by a human trusting and responding to a signal when the system presents one [65]. An increase in the miss rate reduces reliance, while an increase in false alarms reduces compliance [5, 66]. This distinction is important as it leads the human to react to a signal. For example, a compliance-oriented system (i.e., gives warning when there is a malfunction) increases awareness in humans especially when warnings are spaced close to other indicators [67].

Humans may choose to ignore warnings if they experience high rates of false alarms, which is known as the 'cry wolf' effect [68]. This behavior represents humans' mistrust of autonomous systems and induces disuse of these systems [69]. Some studies suggest that false alarms cause greater negative effects on human trust in automation as they divert humans' attention, causing them to monitor unnecessary information [70]. Pervasive false alarms may make humans respond slower or less

frequently to future similar alerts [71, 72]. However, the high false-alarm rate does not appear to negatively impact trust in the context of en route Air Traffic Controller conflict alerts [73]. Indeed, some studies showed contrary results where false alarm-prone systems were more trustworthy than miss-prone systems [74, 75]. In addition, there are studies suggesting that false alarms and misses lead to similar effects on trust [76, 77] or that the effect is dependent on humans' cognitive capabilities [78].

Existing literature shows evident differences in opinions of the effects of misses and false alarms on human trust in automation. In order to resolve these differences, a model for human trust behavior with respect to false alarms and misses is needed. Moreover, the alarm threshold is determined based on the costs associated with each type of error, which means the optimal rate of misses/false alarms varies between systems. Therefore, a model of trust dynamics that connects human trust to autonomous system reliability can help us better understand how reliability-induced trust changes over time. Furthermore, it would allow us to understand how trust recovers with a hit (i.e., system correctly detects the presence of a signal) and/or a correct rejection (i.e., system not alerting the human to an absence of a signal).

Demographic Factors that Influence Trust

Autonomous system reliability and error type are external factors that influence learned trust. Apart from experiences accumulated from past interactions with autonomous systems, human trust behavior is also influenced by demographic factors including culture and gender. This is described as *dispositional trust* and is independent of a specific system or the context of an interaction [12].

Gender differences in trust behavior have been studied thoroughly in economic contexts [35, 79, 80]. Furthermore, some studies have shown gender differences in human-robot interaction (HRI) contexts and technology adoption behavior [36,81,82]. For example, males were more likely to develop trust and positive attitudes toward female robots, while women showed little preference [83]. The attitudes of children

toward humanoid robots are also influenced by gender. Tung showed that girls favored human-like, female robots more than boys [84]. In addition, females perceived highly automated driving systems as significantly less trustworthy than males did [85].

Values and social norms shared by members of a nation that guide people's behaviors and beliefs can be defined as the national culture for each country [86]. These factors also have an influence on the cognitive process of trust formation in humans. Therefore, people from different cultures are likely to use different mechanisms to form trust [87] and show particular trust behavioral intentions [88]. To date, only a few studies have examined the effect of national culture on trust in automation. Rice et al. observed that Americans tended to trust less in automated systems as compared to Indians in the context of "auto-pilots" [89]. In another study, Americans were found to trust autonomous (decision-aid) systems less than Mexicans in a fraud investigation scenario [90]. Trust can also be seen as "the willingness to take risk" [91]. Considering the influence of national culture, Uncertainty Avoidance Index (UAI) defined in Hofstede's six cultural dimensions [86,92] is relevant to the construct of trust. Uncertainty avoidance tendency has been found to be significant in influencing trust in web design attributes [93], mobile commerce [94], information technology infrastructure [95], and in the context of simulated unmanned air vehicle control [96]. The higher the UAI number for a country, the less likely their people will tolerate uncertainty or risk.

In summary, published quantitative dynamic trust models do not explicitly consider a number of different factors, including the nature of human bias toward the system's response criteria (i.e., liberal and conservative), demographics, and false alarms and misses in autonomous systems. Moreover, although literature in the area of multi-agent systems has analyzed the effects of past experiences on future trust level, this effect needs to be modeled for independent human-machine interactions. Therefore, the influence of these factors on human trust *dynamics* remains unexplored. To address these key gaps, we present two experiments that test the trust factors mentioned above and aid us in establishing a dynamic model of human trust.

2.1.3 Experiment 1

The first experiment was designed to understand human trust dynamics induced by the autonomous system performance and to identify the effects of humans' national culture and gender on trust behaviors. The rates of misses and false alarms were controlled, so participants encountered approximately equal numbers of these two error types. In addition, the order of these two error types was randomized within faulty trials. Therefore, the trust behavior induced by a specific error type was neutralized in experiment 1.

Method

Stimuli and procedures. The experiment was conducted online, and each participant accessed the study through a computer interface. Participants were told that the experiment was a simplified simulation of driving a car equipped with an obstacle detection sensor. The sensor was based on an image-recognition algorithm that would detect obstacles on the road in front of the car. During each trial, participant's task was to decide whether or not to trust the algorithm report, based on their previous experience with the algorithm. The instructions informed participants that the image-recognition algorithm used in the sensor was in beta testing.

An experiment session consisted of four initial practice trials followed by 100 trials comprising a sequence of events including stimulus, response, and feedback (see Figure 2.1). There were two stimuli: 'obstacle detected' and 'clear road', each having a 50% probability of occurrence. After receiving the stimulus, participants were asked to determine whether they 'trust' or 'distrust' the report provided by the algorithm. The system then gave feedback to the participants on the correctness of their responses (i.e., 'correct' and 'incorrect'). In order to examine how system reliability influences human trust level, the system was 'reliable' in half of the trials and was 'faulty' in the remaining half. Here reliability is defined as the degree to which the algorithm report can be depended on to be accurate. In reliable trials, the

algorithm correctly identified the road condition. This meant that 'obstacle detected' was a hit, and 'clear road' was a correct rejection. Accordingly, it would be marked as 'correct' if the participant trusted the report, and 'incorrect' if the participant distrusted the report. In faulty trials, there was a 50% probability that the algorithm *incorrectly* identified the road condition. A report of 'obstacle detected' could be a false alarm, and 'clear road' could be a miss (see Figure 2.2). For the participant, this meant that responding 'trust' to a false alarm or a miss would be marked as 'incorrect'. We implemented the 100% accuracy condition for reliable trials because it is the ideal performance a sensor can achieve. On the other hand, 50% accuracy for a binary decision would be a pure random chance. Therefore, it should result in the lowest possible trust level that a human has in the simulated sensor.

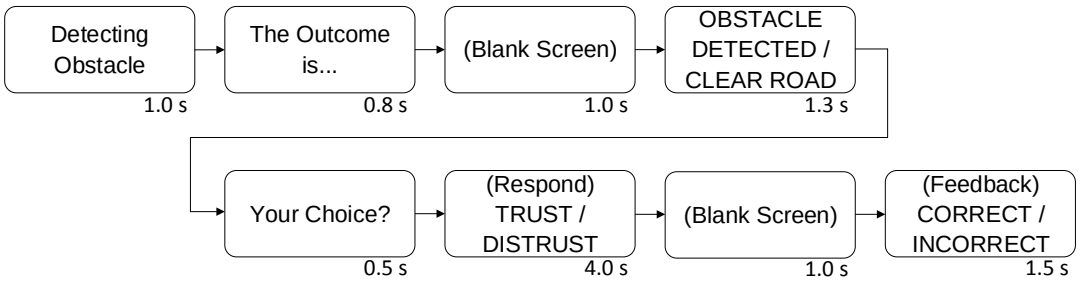


Figure 2.1. Sequence of events in a single trial. The time length marked on the bottom right corner indicates the time interval that the information appeared on the screen.

All of the trials in the study (100 in total) were divided into three phases, called 'databases', as shown in Figure 2.3. There was a 30-second break before the start of each database. Databases 1 and 2 were used to induce responses to constant system reliability—either reliable or faulty. Database 3 was used to excite all possible dynamics of the participants' trust responses by switching the accuracy of the algorithm between reliable and faulty according to a pseudo-random binary sequence (PRBS). Stimuli in each trial were individually randomized for each participant and database.

		Actual Scenario	
		Obstacle Present	**Obstacle Absent**
System Response	**Obstacle Detected**	Hit	False Alarm
	Clear Road	Miss	Correct Rejection

Figure 2.2. The actual scenario and the system response form a 2 × 2 matrix. A system response of 'clear road' in the presence of an obstacle constitutes a miss, and a system response of 'obstacle detected' in the absence of an obstacle constitutes a false alarm.

Participants were randomly assigned to one of two groups which differed in the order of reliable and faulty trials to counterbalance possible ordering effects.

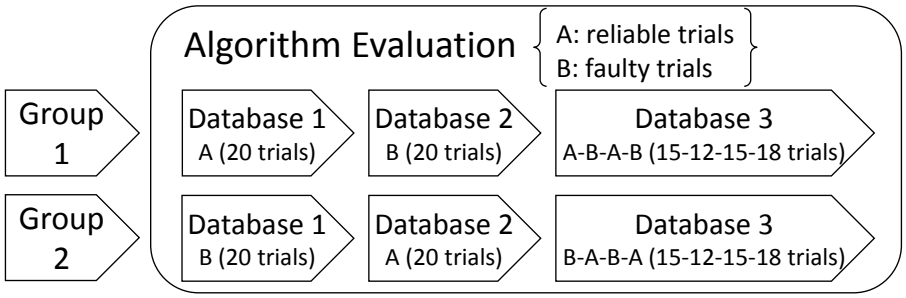


Figure 2.3. Participants were randomly assigned to one of the two groups. The ordering of the three experimental sections (databases), composed of reliable and faulty trials, were counterbalanced across groups.

Participants. A total of 581 individuals (ages 20–73) were recruited using Amazon Mechanical Turk [97] to participate in the study. Among the participants, 340 were males, 235 were females, and six did not provide gender information. These participants were randomly assigned to one of two experimental groups. Participants in groups 1 and 2 were initially faced with reliable trials and faulty trials, respectively. The participants were each paid $0.50 for their participation in the study.

Before starting the study, participants electronically provided their consent. The Institutional Review Board at Purdue University approved the study. We collected participants' demographic information via a post-study survey which included questions about their gender along with the country in which they grew up. The latter is defined as national culture in this study.

Data processing. To pre-process the collected data, we identified and removed outliers from the data set. Each participant completed all 100 trials, but they were allowed to skip a trial if they could not make a decision within the given time frame (4 seconds). We considered excessive "no responses" (i.e., when participants skipped a trial) as well as excessive trust or distrust responses as outliers, determined by the interquartile range (IQR) rule (the 1.5×IQR rule) [98]. As a result, we removed 63 outliers from the dataset (out of 581 participants) which resulted in 518 valid participants.

To investigate the effects of national culture and gender on human trust, we categorized the collected data into four demographic bins; two were based on nationality: United States (US) and India, and two were based on gender: male and female. Ideally, the selected sample would be representative of the population it came from. However, practically it was not possible to have an equal representation of each demographic group in the collected sample. In order to correct this anomaly in the selection probability of each demographic group in the population, the variables of each bin were adjusted using sampling weights such that each group had an equal representation in the sample population [99]. We calculated sampling weights for each demographic group in all of the bins as follows:

$$\text{Sampling weight} = \frac{\text{Population percentage}}{\text{Sample percentage}}. \tag{2.1}$$

Trust model description. For groups 1 and 2, we computed the *probability of trust response* for each trial and across all subjects in each of the groups. This probability is defined as the percentage of people in the group who trust the algorithm report.

At each trial, for calculating this probability, we assume that the response of each participant is like a Bernoulli trial with 'trust' response as success and 'distrust' response as failure. Given that for each trial, the responses of all participants are independent from one another, the random variable X, defined as the number of participants responding 'trust' on a given trial, has a binomial distribution, $B(k, p)$. The parameter k is the total number of participants in the bin and the parameter p, binomial proportion, can be estimated using a normal approximation as $\hat{p} = \frac{k_S}{k}$. Here k_S is the number of successes, i.e., number of trust responses in the given trial across participants. Therefore, at each trial, the probability of trust response can be estimated as $\hat{p}$ for that trial. The range of estimated probabilities was 0.5 to 1, where 0.5 represented low trust (i.e., the report was perceived as random by the participants; therefore, they responded randomly) and 1 represented high trust. These trust probabilities varied as the decision scenario changed with time and represent the trust level for the sample population. Henceforth, the trust probability will be labeled as *trust level* $T(n)$, where $n \in [1, 100]$ is the trial number. Similarly, we calculated the *probability of misses* $M(n)$ and the *probability of false alarms* $F(n)$ for each trial, across all subjects, in groups 1 and 2. For experiment 1, $M(n) = F(n)$ and varied from approximately 0 to 0.25, with 0.25 representing faulty trials leading to negative experience and 0 representing reliable trials leading to positive experience. Therefore, we define *experience* $E(n)$ as a function of $M(n)$ and $F(n)$ given by

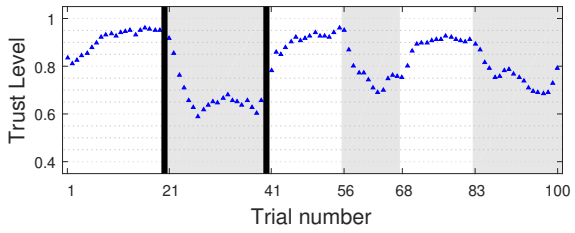
$$E(n) = 1 - [(1 - \beta)M(n) + \beta F(n)]. \tag{2.2}$$

Here, $\beta \in (0, 1)$ is the weighting factor for evaluating the relative effect of misses and false alarms on experience. Beta is the coefficient of the probability of false alarms in the model and thus can be called the *cry-wolf factor*. The higher the value of the cry-wolf factor β, the greater the effect of false alarms on the experience, and the

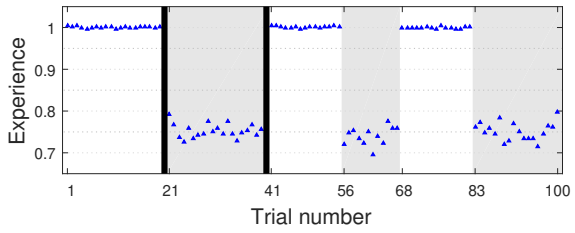
less the effect of misses on the experience. Since, the probability of misses and false alarms was equal for experiment 1 ($M(n) = F(n) = K(n)$), (2.2) reduces to

$$E(n) = 1 - K(n). \tag{2.3}$$

where $K(n)$ is the probability of a miss or a false alarm. Thus we obtain the dynamic variation of trust level $T(n)$ with experience $E(n)$ for all participants as shown in Figure 2.4. In order to reduce noise from the dynamically varying signal $T(n)$, we used the Savitzky-Golay filter with order 3 and a window of size 5 [100].



(a)



(b)

Figure 2.4. The trust level (probability of trust response) and the experience for all participants. The top figure (a) shows the variation of trust level as a function of trial number. The bottom figure (b) shows the variation of experience as a function of trial number. Faulty trials are highlighted in gray, and black lines mark the breaks. Participants showed trust in reliable trials and distrust in faulty trials.

Most of the existing human trust models showed trust to be directly related to experience. Jonker and Treur presented change in trust to be directly proportional to the difference of experience and past trust [33]. However, along with experience, we

identified the significance of *cumulative perception of trust* and *the human's expectations of the autonomous system* in formulating human trust behavior. Therefore, we adapt the model used by Jonker and Treur and introduced two additional terms—*Cumulative Trust* (C_T) and *Expectation Bias* (B_X)—to propose a second order model as shown in (2.4). The states of the model are defined as trust level $T(n)$ and cumulative trust $C_T(n)$, and the input is defined as experience $E(n)$ along with a constant input disturbance called expectation bias B_X.

$$T(n+1) - T(n) = \alpha_e[E(n) - T(n)] \tag{2.4a}$$
$$+ \alpha_c[C_T(n) - T(n)] \tag{2.4b}$$
$$+ \alpha_b[B_X - T(n)] \tag{2.4c}$$
$$C_T(n+1) = [1-\gamma]C_T(n) + \gamma T(n) \tag{2.4d}$$

In the model (2.4), change in trust $T(n+1)-T(n)$ depends linearly on three terms: $E(n) - T(n)$ (2.4a), $C_T(n) - T(n)$ (2.4b), and $B_X - T(n)$ (2.4c), where each term is bounded between -1 and 1. We call the parameters α_e, α_c, and α_b the *experience rate factor*, *cumulative rate factor*, and *bias rate factor*, respectively, since they control the rate by which each individual difference affects the predicted trust level. Details on the estimation of these parameters are described in the *Parameter Estimation* subsection.

As shown in (2.4d), we define cumulative trust C_T as an exponentially weighted moving average of past trust level. Cumulative trust incorporates the learned trust in the model using a weighted history of past trust levels. A higher value of the parameter γ discounts older trust levels faster, and thus γ can be called the *trust discounting factor*. The expectation bias B_X accounts for a human's expectation of a particular interaction with an autonomous system. This is modeled as an input disturbance which remains constant during an interaction. The state $T(n)$ represents a probability, and the state $C_T(n)$ represents an exponentially weighted moving average of probability; therefore both belong to $[0, 1]$. $B_X(n)$ must belong to $[0, 1]$ so that

$T(n+1)$ remains bounded within $[0,1]$ in (2.4a). Moreover, $E(n)$ belongs to $[0,1]$, based on (2.2).

The linearity of the proposed model allows us to represent the model in state space form as

$$\begin{aligned} x(n+1) &= \begin{bmatrix} 1-\alpha & \alpha_c \\ \gamma & 1-\gamma \end{bmatrix} x(n) + \begin{bmatrix} \alpha_e \\ 0 \end{bmatrix} u(n) + \begin{bmatrix} \alpha_b \\ 0 \end{bmatrix} d(n) \\ y(n) &= \begin{bmatrix} 1 & 0 \end{bmatrix} x(n) \end{aligned} \tag{2.5}$$

where $x = \begin{bmatrix} T & C_T \end{bmatrix}^T$, $u = E$, $d = B_X$, and $\alpha = \alpha_e + \alpha_c + \alpha_b$. The linearity of the model also simplifies analysis of the trust dynamics as well as potential synthesis of model-based control algorithms for improved human-machine interactions.

Proposition 1 *The linear state-space model given in* (2.5) *is stable if the parameters* α_e, α_c, α_b, α, *and* γ *belong to* $(0,1)$.

Proof The eigenvalues of the proposed discrete model are given by

$$\lambda_{1,2} = 1 - \frac{\alpha+\gamma}{2} \pm \sqrt{\frac{\alpha-\gamma}{2} + \alpha_c\gamma} \tag{2.6}$$

and must lie inside the unit circle, i.e., $|\lambda_{1,2}| < 1$ to guarantee asymptotic stability. Therefore, it is sufficient to prove that

$$\lambda_1 = 1 - \frac{\alpha+\gamma}{2} - \sqrt{\left(\frac{\alpha-\gamma}{2}\right)^2 + \alpha_c\gamma} > -1 \ , \tag{2.7a}$$

$$\lambda_2 = 1 - \frac{\alpha+\gamma}{2} + \sqrt{\left(\frac{\alpha-\gamma}{2}\right)^2 + \alpha_c\gamma} < 1 \ . \tag{2.7b}$$

By rearranging and squaring both sides, (2.7a) and (2.7b) can be reduced to show that $\forall\ \gamma \in (0,1)$,

$$2 > \alpha + \alpha_c \quad \text{and} \tag{2.8a}$$

$$0 < \alpha_e + \alpha_b\ . \tag{2.8b}$$

Equations (2.8a) and (2.8b) are satisfied if $\alpha_e, \alpha_c, \alpha_b \in (0,1)$ and $\alpha \in (0,1)$. Therefore, the trust model (2.5) is asymptotically stable. ■

Remark 1 *The physical interpretation of these bounds on the parameters can be obtained by closer examination of* (2.4)*. The parameters* α_e*,* α_c*, and* α_b *are weighting factors for each of the terms and should be less than 1 so that the trust level remains stable. The variable* γ *is an exponential weighting factor that belongs to* $(0,1)$*. Additionally,* $\alpha = \alpha_e + \alpha_c + \alpha_b$ *belongs to* $(0,1)$*. This ensures that the net coefficient of the term* $T(n)$ *for calculating* $T(n+1)$*, i.e.,* $1-\alpha$*, belongs to* $(0,1)$ *and is not negative. Consequently, a higher previous trust level will have a positive influence on current trust level.*

Proposition 2 *The steady-state values of trust,* T_{ss}*, and cumulative trust,* C_{Tss}*, for a stable system given by* (2.5) *are a weighted average of steady-state experience* E_{ss} *and expectation bias* B_X*. The weights are proportional to* α_e *and* α_b*.*

Proof By substituting $x(n+1)$ and $x(n)$ with $x_{ss} = [T_{ss} \quad C_{Tss}]^\top$ and $u(n)$ with $u_{ss} = E_{ss}$, in (2.5), we can solve for the steady-state values T_{ss} and C_{Tss} as follows:

$$T_{ss} = C_{Tss} = \frac{\alpha_e}{\alpha_e + \alpha_b} E_{ss} + \frac{\alpha_b}{\alpha_e + \alpha_b} B_X\ . \tag{2.9}$$

Here the subscript $\bullet_{ss}$ represents the steady-state value of the variable. ■

Remark 2 *Consider the case when* $E_{ss} = 1$ *which indicates that the system interacting with the human is consistently accurate. If the expectation bias is less than 1*

($B_X < 1$), the steady-state trust level T_{ss} of the human will be less than 1. Alternatively, consider the case when $E_{ss} = 0$, which indicates that the system interacting with the human is consistently faulty. If $B_X > 0$, the steady-state trust level T_{ss} will also be greater than 0. Therefore, the inclusion of human bias in the proposed model enables us to characterize this important effect on human trust level.

Parameter Estimation. For estimating the optimal set of model parameters, we used a nonlinear least squares estimation function *nlgreyest* from MATLAB 2016a. We identified the parameters using 1) the data of all participants and 2) the data in each of the four demographic bins. Each dataset consisted of data from each of the three 'databases' in both group 1 (in which participants were initially faced with reliable trials) and group 2 (in which participants were initially faced with faulty trials). It is well known that the quality of any empirical parameter estimation is dependent on the data itself. A sample of human subject data cannot completely represent the human population, and the derived inferences may vary based on the selected sample. Therefore, in order to verify the robustness of the parameter estimation relative to the sample selection, we iterated the estimation 1000 times, with each iteration using a new randomly selected subset of data representing 90% of the total dataset for all participants and each demographic bin. There was less than 2.5% error in the estimated parameter values caused by the variation in sample selection for a 95% confidence interval (CI) (see Table 2.1), signifying a robust estimation.

Table 2.1.
Estimated mean parameter values with 95% CI for all participants and each demographic bin

Bin	Experience rate factor α_e	Cumulative rate factor α_c	Bias rate factor α_b	Trust discounting factor γ	Fit% Grp 1	Fit% Grp 2
All	0.2169 ± 0.0007	0.0755 ± 0.0005	0.0428 ± 0.0004	0.1148 ± 0.0012	95.71	92.56
US	0.2157 ± 0.0007	0.0635 ± 0.0007	0.0394 ± 0.0004	0.1270 ± 0.0029	94.59	87.59
India	0.2177 ± 0.0010	0.0996 ± 0.0011	0.0515 ± 0.0007	0.0942 ± 0.0008	91.97	90.14
Female	0.2277 ± 0.0009	0.0783 ± 0.0007	0.0373 ± 0.0005	0.1042 ± 0.0017	91.48	89.12
Male	0.2085 ± 0.0009	0.0817 ± 0.0007	0.0491 ± 0.0005	0.1375 ± 0.0018	93.93	89.17

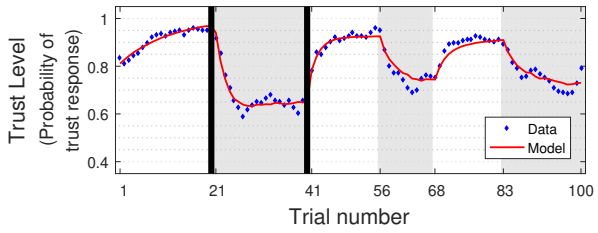
Results

In order to verify whether our proposed model of trust level is valid, we estimated the model parameters for a general population, which included all 518 valid participants in our experiment. The fit between the trust model and the experimental data is shown in Figure 2.5. Table 2.1 shows the optimal parameter values and the goodness of fit between the data and the model calculated using R-squared. The goodness of fit was 95.71% and 92.56% for all participants in groups 1 and 2, respectively. Note that all of the estimated parameter values satisfy the stability criteria defined in Proposition 1.

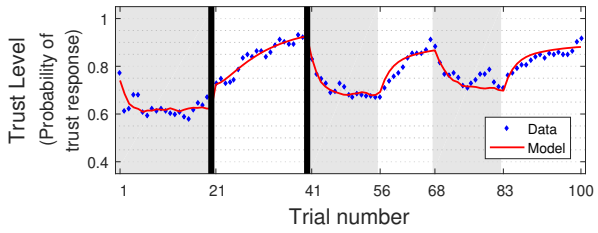
We observed that participants from different demographic groups required different amounts of time to adapt to changes in the system performance and attained different steady-state trust levels. In order to analyze these differences, we simulated the step response of each parameterized model. A sample step response for all participants with expectation bias $B_X = 0$ is shown in Figure 2.6. The calculated rise time for the step response (shown in Table 2.2) is an indicator of the rate of change of the trust dynamics. Rise time is defined as the time required for the response to increase from 10% to 90% of its final value. Therefore, a longer rise time implies slower trust dynamics.

Table 2.2.
Rise times (in number of trials) for step responses calculated using the estimated parameter values for all participants and each demographic bin

Bin	Rise Time (Number of Trials)	
	T	C_T
All	15	27
US	13	24
India	20	34
Female	13	27
Male	15	23



(a)



(b)

Figure 2.5. Participants' trust level (blue dots) and the prediction (red curve) based on past behavioral responses and the experience of all participants. Subfigure (a) corresponds to group 1 participants with $R^2 = 95.74\%$ and subfigure (b) corresponds to group 2 participants with $R^2 = 92.53\%$. Faulty trials are highlighted in gray, and black lines mark the breaks between databases.

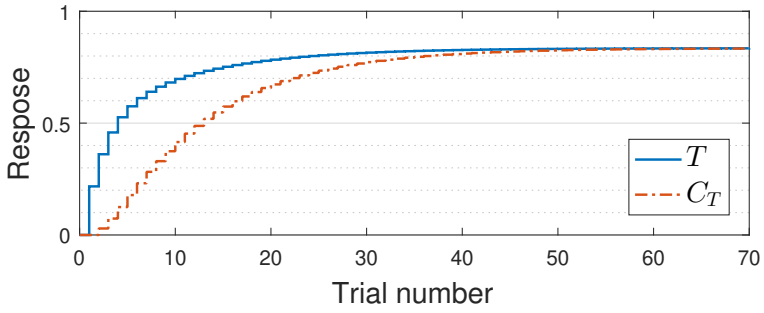
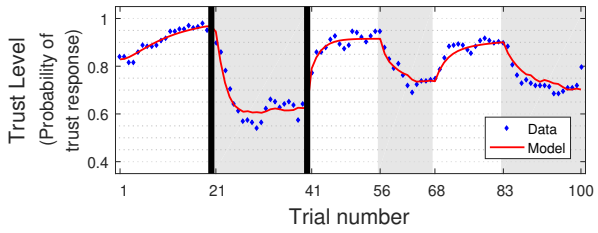
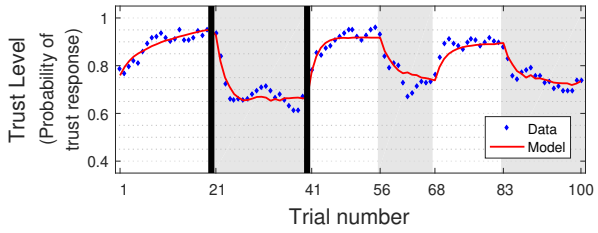


Figure 2.6. Step response of the trust model with expectation bias $B_X = 0$ for all participants.

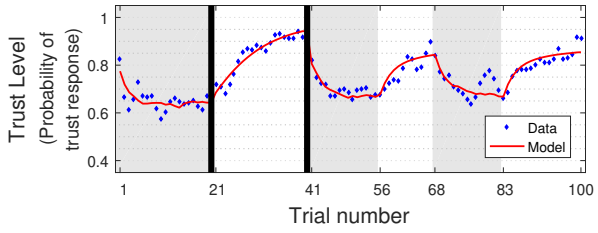
Figure 2.7 shows the experimentally obtained trust level and the predicted trust level of participants grouped by their national culture. Upon visual inspection, US



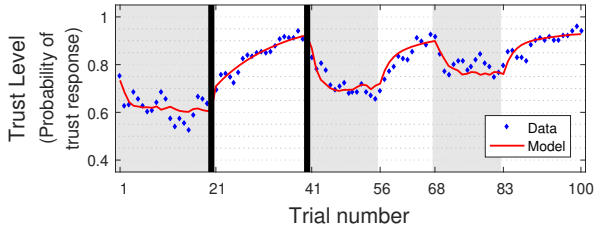
(a)



(b)



(c)



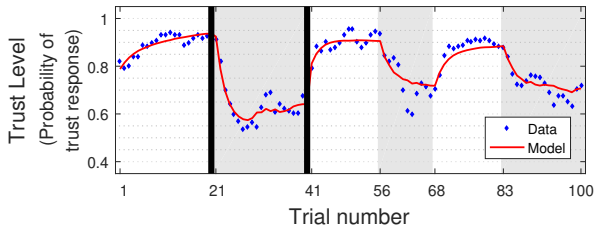
(d)

Figure 2.7. Participants grouped by national culture. Blue dots are the reported trust level while the red curve is the prediction from model. Subfigure (a) corresponds to US group 1 participants with $R^2 = 94.51\%$ and subfigure (b) corresponds to Indian group 1 participants with $R^2 = 92.00\%$. Subfigure (c) corresponds to US group 2 participants with $R^2 = 87.56\%$ and subfigure (d) corresponds to Indian group 2 participants with $R^2 = 90.08\%$. Faulty trials are highlighted in gray, and black lines mark the breaks between databases.

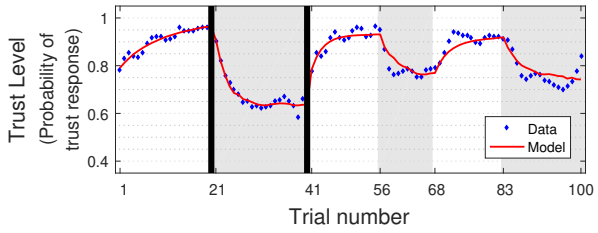
participants trusted the system report less during the trials in database 3, than during trials in databases 1 and 2, in which the accuracy of the algorithm was switched between reliable and faulty; see Figure 2.7(a) and 2.7(c). Moreover, in response to changes in system reliability, the trust level of US participants changed at a faster rate, and approached an overall lower level, than that of Indian participants. These observations are supported by the calculated rise time of the models (Table 2.2). The rise time of the state T for Indian participants is 53.8% higher than that of US participants. This implies that Indian participants' trust level increased or decreased more slowly than that of US participants after the system performance changes. Additionally, the rise time of the state C_T for US participants is 29.4% shorter than that of Indian participants, which implies that their cumulative trust changed relatively faster. This observation can also be attributed to the trust discounting factor γ, which is 34.8% larger for US participants, indicating that US participants relied on their recent trust level and experience more as compared to Indian participants.

Figure 2.8 shows the experimentally obtained trust level and the prediction of participants grouped by their gender. The plots show that male participants exhibited greater trust in the system than female participants, especially when the system did not perform well (see Figure 2.8(b) and 2.8(d)). On the other hand, the trust level of female participants changed more rapidly than that of male participants. Similarly, when comparing the step responses, the rise time of state T for male participants is 15.4% longer than that of female participants, implying that the trust level of male participants changed more slowly than that of female participants. Furthermore, the rise time of the state C_T for male participants is 14.8% shorter than that of female participants, which implies that their cumulative trust changed relatively faster. This observation can also be attributed to the trust discounting factor γ, which is 32.0% larger for male participants, indicating that they relied on their recent trust level more as compared to female participants.

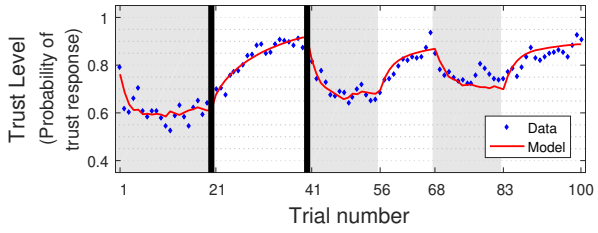
Based upon the high fit percentages achieved between the model and experimental data after parameter estimation, these results suggest that human trust in



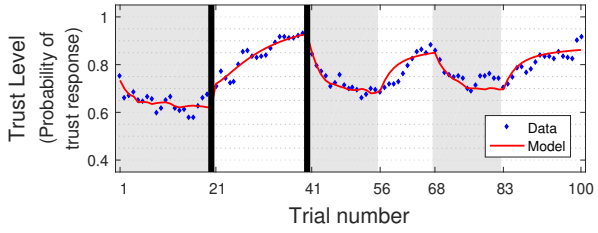
(a)



(b)



(c)



(d)

Figure 2.8. Participants grouped by gender. Blue dots are the reported trust level while the red curve is the prediction from model. Subfigure (a) corresponds to female group 1 participants with $R^2 = 91.57\%$ and subfigure (b) corresponds to male group 1 participants with $R^2 = 93.98\%$. Subfigure (c) corresponds to female group 2 participants with $R^2 = 88.94\%$ and subfigure (d) corresponds to male group 2 participants with $R^2 = 89.22\%$. Faulty trials are highlighted in gray and black lines mark the breaks between databases.

autonomous systems can be modeled as a function of their experience (which varies with system performance), cumulative trust, and expectation bias. Moreover, the estimated model parameters capture the effects of national culture and gender on trust behaviors.

2.1.4 Experiment 2

As an extension of experiment 1, we designed experiment 2 to conduct an in-depth study on the effects of misses and false alarms on participants' trust levels. In this experiment, we present participants with trials containing 100% of misses and 100% of false alarms, unlike the 50-50 split used in experiment 1 (see Figure 2.9).

Method

We followed the same methodologies from experiment 1 in terms of data collection, data processing, and modeling. We revised the stimuli to elicit trust reactions in response to misses and false alarms and analyzed the resulting data that was collected. We then expanded the general trust model to incorporate the effects of misses and false alarms.

Stimuli and procedures. In comparison to experiment 1, the only additional factor incorporated into experiment 2 was the error type during faulty trials. More specifically, we manipulated the probability of misses and false alarms in faulty trials. In experiment 1, a system error was equally probable to be a miss or a false alarm in each faulty trial. In experiment 2, we examined the following three conditions: 1) an error was always a miss in a session of faulty trials; 2) an error was always a false alarm in a session of faulty trials; and 3) an error was equally probable to be a miss or a false alarm in a session of faulty trials. Figure 2.10 shows the condition and trial orders in each database. Participants were randomly assigned to one of two groups in the interest of testing whether the experience of misses or false alarms affects the other.

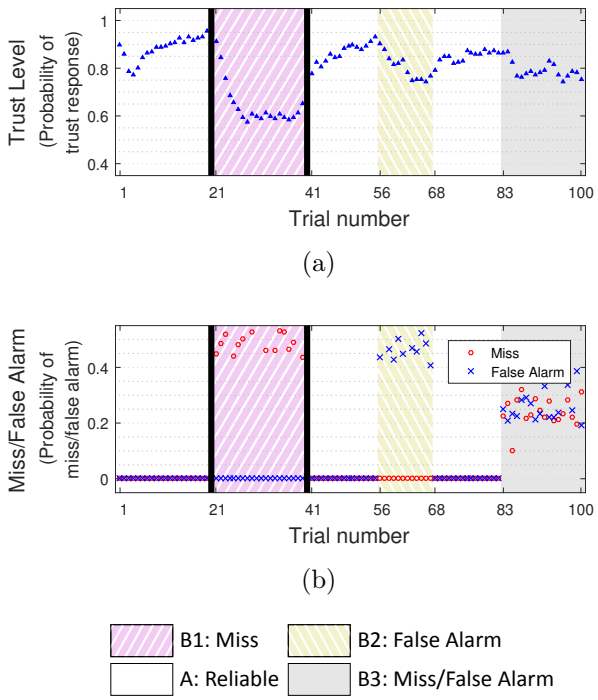


Figure 2.9. The trust level (probability of trust response) for all participants and the probability of misses/false alarms that affect the experience. The top figure (a) shows the variation of trust level as a function of trial number. The bottom figure (b) shows the variation of misses/false alarms as a function of trial number. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks. Participants showed trust in reliable trials and distrust in faulty trials.

Participants. A total of 333 individuals (ages 19–74) participated in experiment 2. Among the participants, 171 were males, 158 were females, and four did not provide gender information. These participants were randomly assigned to one of two experimental groups. The recruitment procedure and the survey used to collect demographic information were the same as in experiment 1.

Data processing. We used the interquartile range (IQR) rule as introduced in experiment 1 to identify and remove outliers. The procedure resulted in 293 valid data sets (out of a total of 333 participants) to be analyzed.

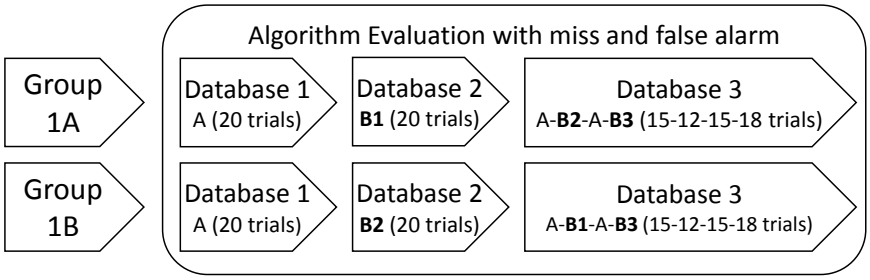


Figure 2.10. Participants were randomly assigned to one of the two groups. The system reliability was varied between databases and groups. *A* consisted of reliable trials (miss = 0%, false alarm = 0%); *B1* consisted of faulty trials with misses (miss = 50%, false alarm = 0%); *B2* consisted of faulty trials with false alarms (miss = 0%, false alarm = 50%); *B3* consisted of faulty trials with both misses and false alarms (miss = 25%, false alarm = 25%)

.

Parameter Estimation. Using the data collected in experiment 2, we estimated the *cry-wolf factor* β by setting all other factors (*experience rate factor* α_e, *cumulative rate factor* α_c, *bias rate factor* α_b, and *trust discounting factor* γ) to the values estimated in experiment 1. The robustness of the estimated value of β was verified by 1000 iterative estimations, with each iteration using a new randomly selected subset of data representing 90% of the total dataset for all participants and each demographic bin. The errors caused by the variation in sample selection for a 95% confidence interval (CI) were less than 2.5%. Table 2.3 shows the parameter values and the goodness of fit.

Results

We first investigated whether the system error type (i.e., miss and false alarm) affects the trust dynamics of the general population, which included all 293 valid participants in the experiment. Figure 2.11 shows the experimental trust level compared against the model. Participants responded differently to misses and false alarms, and

Table 2.3.
Estimated mean parameter values with 95% CI for the cry-wolf factor β for all participants and each demographic bin

Bin	Cry-wolf factor β	Fit% Grp 1	Fit% Grp 2
All	0.3956 ± 0.0012	91.7593	91.2061
US	0.3209 ± 0.0016	90.6699	87.4562
India	0.4758 ± 0.0019	82.4059	87.1428
Female	0.4276 ± 0.0015	83.8733	80.2373
Male	0.3623 ± 0.0018	89.8757	90.8326

in some cases, the experience of one error type affected later responses to the other error type. The proposed trust model was able to predict the trust dynamics while taking into account the rate of misses and false alarms. The goodness of fit was measured using the R-squared value of the data; the result was 91.76% and 91.20% for all participants in groups 1 and 2, respectively.

The results suggest an interaction effect between risk-taking behavior and demographic factors on trust. Figure 2.12 shows the experimentally obtained trust level and model predictions for participants grouped by their national culture. US participants trusted less than Indian participants when encountering system misses (see Figure 2.12(a) and 2.12(b)). Moreover, US participants trusted less in miss-prone trials than false alarm-prone trials regardless of whether they encountered false alarms first or not (see Figure 2.12(a) and 2.12(c)). By contrast, Indian participants trusted less in miss-prone trials than false alarm-prone trials only when they encountered misses first (see Figure 2.12(b)); their trust level in miss-prone trials decreased less if they encountered system false alarms prior to misses (see Figure 2.12(d)). The cry-wolf factor β of the model is 48.3% larger for Indian participants than that of US participants. The larger the value of β, the weaker the negative effect of misses on trust, indicating that misses have a stronger negative effect on trust for US participants as compared with Indian participants.

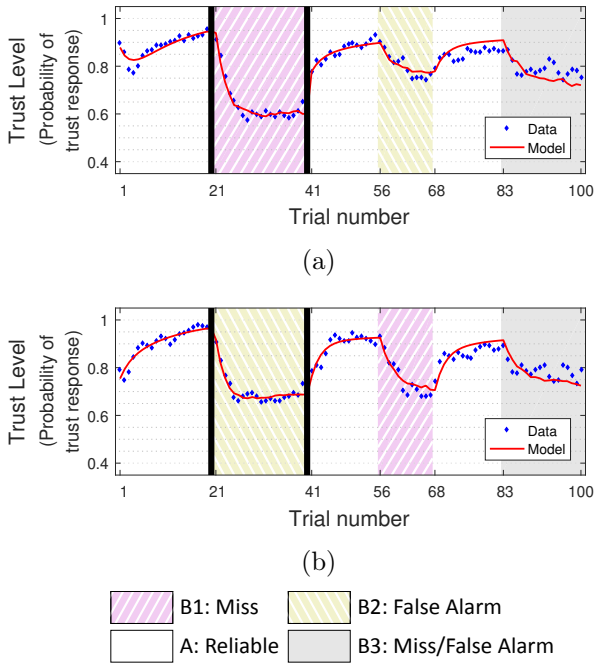


Figure 2.11. Participants' trust level (blue dots) and the prediction (red curve) based on past behavioral responses and the experience of all participants. Subfigure (a) corresponds to group 1A participants with $R^2 = 91.83\%$ and subfigure (b) corresponds to group 1B participants with $R^2 = 91.25\%$. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks between databases.

We also observed gender differences in response to system misses and false alarms. Figure 2.13 shows the experimental trust level and the prediction of participants grouped by their gender. Male participants trusted less in miss-prone trials than female participants if they had not encountered system false alarms first (compare Figure 2.13(a) and 2.13(b)). On the other hand, if participants encountered false alarms first, females reached a lower trust level than males (compare Figure 2.13(c) and 2.13(d)). In general, male participants were more sensitive to system misses. The cry-wolf factor β of the trust model supports this observation; β is 18.0% larger for

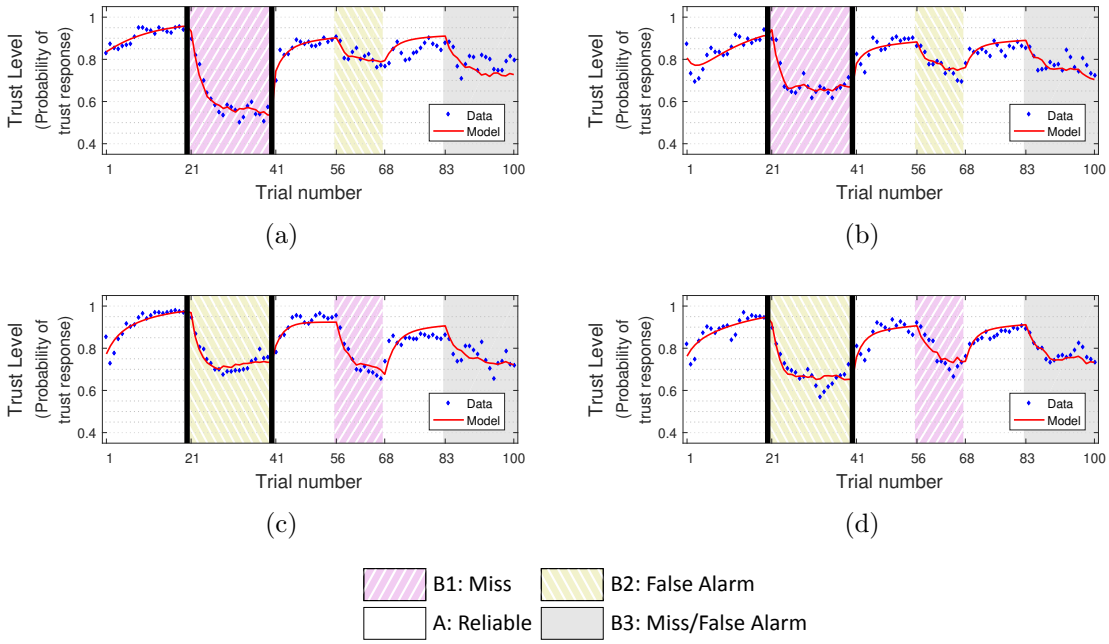


Figure 2.12. Participants grouped by national culture. Blue dots are the reported trust level while the red curve is the prediction from model. Subfigure (a) corresponds to US group 1A participants with $R^2 = 90.67\%$ and subfigure (b) corresponds to Indian group 1A participants with $R^2 = 82.41\%$. Subfigure (c) corresponds to US group 1B participants with $R^2 = 87.46\%$ and subfigure (d) corresponds to Indian group 1B participants with $R^2 = 87.14\%$. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks between databases.

female participants than male participants, which implies that misses have a stronger negative effect on trust for male participants as compared with female participants.

2.1.5 Discussion

Here, we provide a more in-depth discussion of the main results of the two experiments. The two experiments presented in this section elicited the *variation* of a human's trust response to system reliability. Participants attained a high trust level

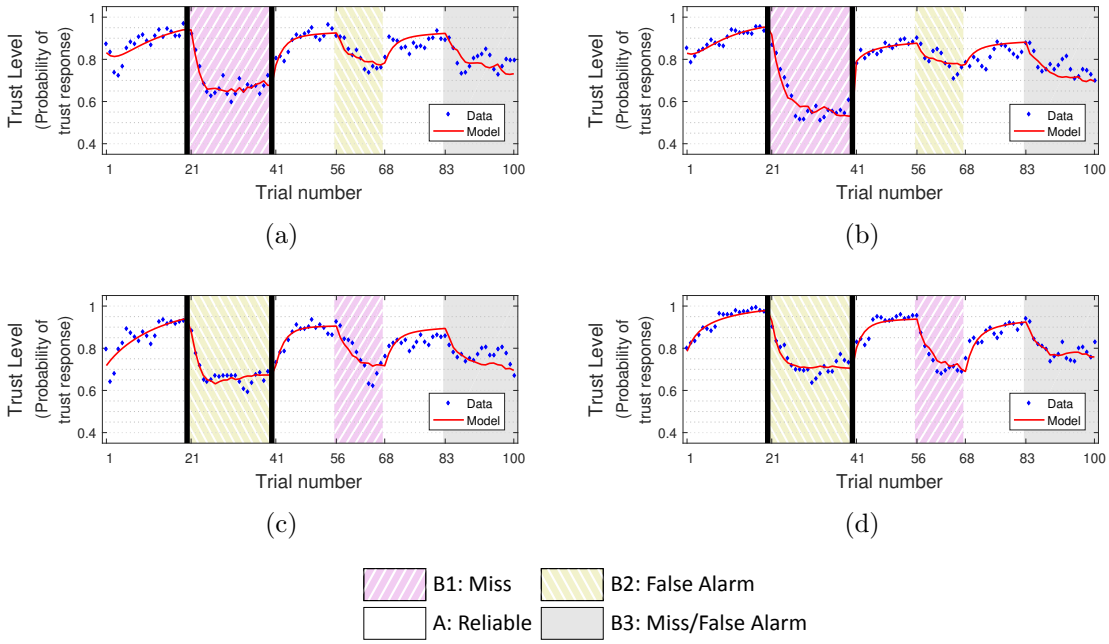


Figure 2.13. Participants grouped by gender. Blue dots are the reported trust level while the red curve is the prediction from model. Subfigure (a) corresponds to female group 1A participants with $R^2 = 93.87\%$ and subfigure (b) corresponds to male group 1A participants with $R^2 = 89.88\%$. Subfigure (c) corresponds to female group 1B participants with $R^2 = 80.24\%$ and subfigure (d) corresponds to male group 1B participants with $R^2 = 90.83\%$. Faulty trials consisting of misses are highlighted in pink, and trials with false alarms are highlighted in yellow. Faulty trials highlighted in gray consist of half misses and half false alarms. Black lines mark the breaks between databases.

in reliable trials and a low trust level in faulty trials; this was achieved without training the participants or providing them with specific information (e.g., a game rule or background stories). The trust dynamics were modeled based on *past behavioral responses of the human*, *human trust bias*, and the system reliability. The system reliability was further described by *the rate of misses and false alarms*. The model was verified using the collected human subject data that accounted for ordering effects with respect to system reliability. In other words, the prediction capability of the model was consistent for both groups 1 and 2. Thus, the model can describe

human trust irrespective of the initial condition of the system reliability. Moreover, the interaction between the human and machine was the most significant factor in temporal variations in trust level. Therefore, the developed study is effective for modeling dynamic human trust behavior in HMI contexts.

The proposed study design induced trust dynamics by manipulating multiple transitions between positive and negative experiences. We observed that it took approximately eight to ten trials for participants to establish a new trust level. Moreover, in some cases (e.g., Figure 2.13(a)) the trust response still increased or decreased near this newly attained level in both reliable and faulty trials. This finding was contrary to Jonker et al. who asserted that "after a negative experience an agent will trust less or the same amount, but never more" [58]. Jonker's study was composed of only two sets of five trials, each with one transition in between. However, we found this to be less than the required number of trials to reach a new trust level.

The aggregated trust response and the trust model enhanced our understanding of dispositional trust and learned trust in autonomous systems. Participants from the US exhibited a lower trust level than Indian participants. This is consistent with the findings from Huerta et al. and Rice et al. that Americans trust autonomous systems less than Mexicans and Indians, respectively [89,90]. Moreover, system misses induced stronger distrust in US participants than in Indian participants, suggesting that US participants are less willing to take risks. This agrees with the smaller Uncertainty Avoidance Index of Indian culture as compared to that of US culture (40 vs. 46) [38], where the literature demonstrated that humans from higher uncertainty avoidance cultures are less likely to trust or implement new technology [94, 95].

Regarding gender, male participants appeared to trust more than female participants, especially when the system was not reliable. This is supported by Feldhütter et al. [85], but is contrary to the findings of Haselhuhn et al. [101] which showed that women's trust decreases less than men after transgressions as they prefer to maintain interpersonal relationships. These results highlight that the dynamics of human trust behavior in HMI contexts is different from interpersonal trust behavior between hu-

mans, thus creating a need for human trust models in HMI contexts. Additionally, the variation in trust responses of female participants was noticeably higher than that of male participants. This variation indicates that the female participants have diverse perceptions of autonomous systems and therefore, other factors such as personality and expertise should be investigated in future studies. Finally, there were gender differences in the responses to misses and false alarms as discussed in Section 2.1.4. Along with the observations of US and Indian participants, demographic effects can partially explain the inconsistency between previously published results on the effects of system error type on human trust.

We identified the significance of cumulative trust and expectation bias through experiments that elicited multiple dynamic transitions in human trust, and then incorporated these two variables in the proposed linear model. In addition to proposing a general trust model structure, we characterized the effects of both dispositional and learned trust factors, specifically national culture, gender and system error type, using estimated model parameters. We also characterized the effects of misses and false alarms on the dynamics of human trust behavior and compared differences between demographics. While the proposed model is representative of a population of individuals rather than trained to a specific human, such a model could be used to design machines that are required to interact with unspecified users grouped by demographics.

One limitation of this study is that a computer-based interface system was used in the experiment, and therefore, the ecological validity could be improved by conducting experiments in real-life settings. The model could also be generalized for use in a wider range of domains by expanding the definition of experience to incorporate other significant factors beyond the probability of misses and false alarms, such as system transparency or the level of automation. In the next section, we will discuss the use of human psychophysiological measurements, specifically EEG and GSR, for estimation of human trust in real time.

2.2 Estimating Human Trust using Psychophysiological Measurements

The contents of this section were previously published by Akash, Hu, Jain, and Reid in *ACM Transactions on Interactive Intelligent Systems* [102] and are reported here with minor modifications.

2.2.1 Introduction

In the last section, we presented a dynamic model of human trust that captures the effects of automation reliability and error type (miss or false alarm). Other researchers have also attempted to predict human trust using dynamic models that rely on the experience and/or self-reported behavior of humans [33, 34]. However, it is not practical to retrieve human self-reported behavior continuously for use in a feedback control algorithm. Other than dynamic quantitative models, an alternative is the use of psychophysiological signals to estimate trust level [39]. While these measurements have been correlated to human trust level [40, 41], they have not been studied in the context of real-time trust sensing.

In this section we present a human trust sensor model based upon real-time psychophysiological measurements, primarily galvanic skin response (GSR) and electroencephalography (EEG). The model is based upon data collected through a human subject study and the use of classification algorithms to estimate human trust level using psychophysiological data. The proposed methodology for real-time sensing of human trust level will enable the development of a machine algorithm aimed at improving interactions between humans and machines.

This section is organized as follows. In Section 2.2.2 we introduce related work in human-machine interaction, psychophysiological measurements, and their applications in trust sensing. We then describe the experimental study and data acquisition in Section 2.2.3. The data pre-processing technique for noise removal is presented in Section 2.2.4 along with EEG and GSR feature extraction. In Section 2.2.5, we demonstrate a 2-step feature selection process to obtain a concise and optimal feature

set. The selected features are then used for training Quadratic Discriminant Analysis classifiers in Section 2.2.6, followed by model validation.

2.2.2 Background

There are few psychophysiological measurements that have been studied in the context of human trust. We focus here on electroencephalography (EEG) and galvanic skin response (GSR) which are both noninvasive and whose measurements can be collected and processed in real-time. EEG is an electrophysiological measurement technique that captures the cortical activity of the brain [42]. These brain activities exhibit changes in human thoughts, actions, and emotions. Brain-Computer Interface (BCI) technology utilizes EEG to design interfaces that enable a computer or an electronic device to understand a human's commands [103, 104]. The most extensive approach used to identify EEG patterns in BCI design includes feature selection and classification algorithms as they typically provide good accuracy [105].

Some researchers have studied trust via EEG measurements, but only with event-related potentials (ERPs). ERPs measure brain activity in response to a specific event. An ERP is determined by averaging repeated EEG responses over many trials to eliminate random brain activity [42]. Boudreau *et al.* found a difference in peak amplitudes of ERP components in human subjects while they participated in a coin toss experiment that stimulated trust and distrust [40]. Long *et al.* further studied ERP waveforms with feedback stimuli based on a modified form of the coin toss experiment [41]. The decision-making in the "trust game" [106] has been used to examine human-human trust level. Although ERPs can show how the brain functionally responds to a stimulus, they are event-triggered. It is difficult to identify triggers during the course of an actual human-machine interaction, thereby rendering ERPs impractical for real-time trust level sensing.

GSR is a classical psychophysiological signal that captures arousal based upon the conductivity of the surface of the skin. It is not under conscious control but is instead

modulated by the sympathetic nervous system. GSR has also been used in measuring stress, anxiety, and cognitive load [107, 108]. Researchers have examined GSR in correlation with human trust level. Khawaji *et al.* found that average GSR values, and average GSR peak values, are significantly affected by both trust and cognitive load in the text-chat environment [43]. However, the use of GSR for *estimating* trust has not been explored and was noted as an area worth studying [39]. With respect to both GSR and EEG, a fundamental gap remains in determining a static model that not only estimates human trust level using these psychophysiological signals but that is also suitable for real-time implementation.

2.2.3 Methods and Procedures

In this section we describe a human subject study that we conducted to identify psychophysiological features that are significantly correlated to human trust in intelligent systems, and to build a trust sensor model accordingly. The experiment consisted of a simple HMI context that could elicit human trust dynamics in a simulated autonomous system. Our study used a within-subjects design wherein both behavioral and psychophysiological data were collected and analyzed. We then used the data to build an empirical model of human trust through a process involving feature extraction, feature selection, and model training, that is described in Sections 2.2.4, 2.2.5, and 2.2.6, respectively. Figure 2.14 summarizes the modeling framework.

Participants

Participants were recruited using fliers and email lists. All participants were compensated at a rate of $15/hr. The sample included forty-eight adults between 18 and 46 years of age (mean: 25.0 years old, standard deviation: 6.9 years old) from West Lafayette, Indiana (USA). Of the forty-eight adults, sixteen were females and thirty-two were males. All participants were healthy and one was left-handed. The group of participants were diverse with respect to their age, professional field, and cultural

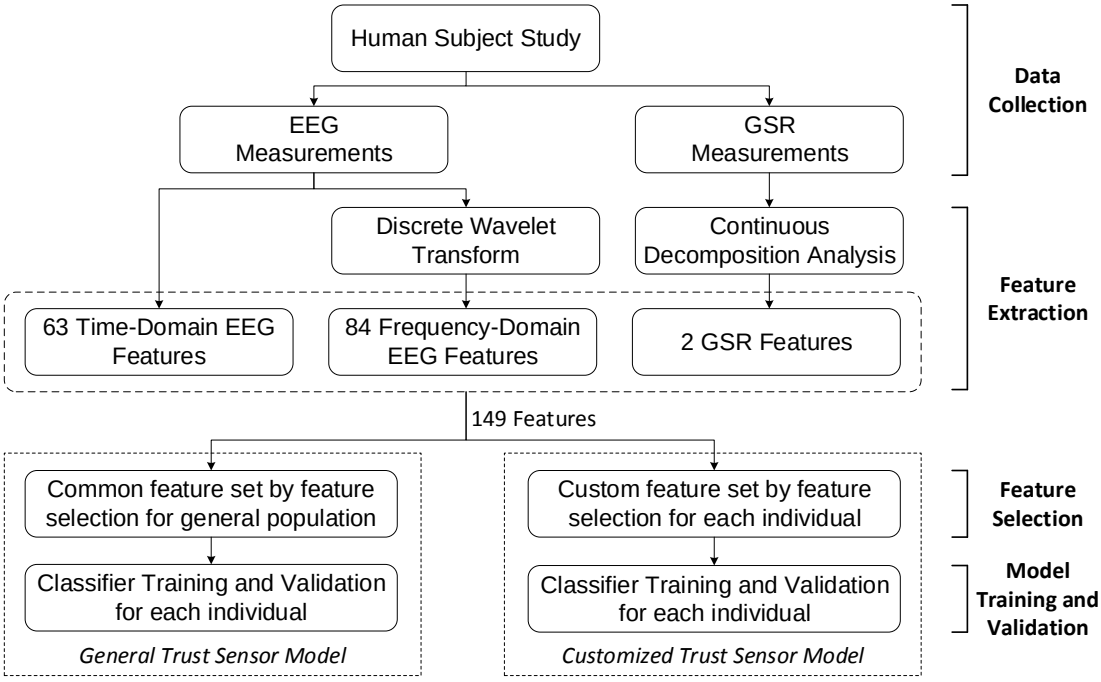


Figure 2.14. The framework of the proposed study. The key steps include data collection from human subject studies, feature extraction, feature selection, model training, and model validation.

background (i.e., nationality). The Institutional Review Board at Purdue University approved the study.

EEG and GSR Recording

EEG: The participant's brain waves were measured using a B-Alert X-10 9-channel EEG device (Advance Brain Monitoring, CA, USA), at a frequency of 256 Hz from 9 scalp sites (Fz, F3, F4, Cz, C3, C4, POz, P3, and P4 based on the 10-20 system). All EEG channels were referenced to the mean of the left and right mastoids. The surface of all sensor sites was cleaned with 70% isopropyl alcohol. Conductive electrode cream (Kustomer Kinetics, CA, USA) was then applied to each electrode including the reference. The contact impedance between electrodes and skin was kept to a value

less than 40 kΩ. The EEG signal was recorded via iMotions (iMotions, Inc., MA, USA) on a Windows 7 platform with Bluetooth connection.

GSR: The skin conductance was measured from the proximal phalanges of the index and the middle fingers of the non-dominant hand (i.e., on the left hand for 43 out of 44 participants) at a frequency of 52 Hz via the Shimmer3 GSR+ Unit (Shimmer, MA, USA). Locations for attaching Ag/AgCl electrodes (Lafayette Instrument, IN, USA) were prepared with 70% isopropyl alcohol. The participants were asked to keep their hands steady on the desk to minimize the influence of movement on the measured signals. The environment temperature was controlled at 72-74°F to minimize the effect of temperature. The GSR signal was also recorded via iMotions so that it would be synchronized with the recorded EEG signals using the common system-timestamps between these two signals.

Experimental Procedure

After the participants read and signed the informed consent, they were equipped with the EEG headset and the GSR sensor as shown in Figure 2.15. All participants finished a 9-minute EEG baseline task provided by Advanced Brain Monitoring and were then instructed to interact with our custom-designed computer-based simulation. Participants were told that they would be driving a car equipped with an image–based obstacle detection sensor. The sensor would detect obstacles on the road in front of the car, and the participant would need to repeatedly evaluate the algorithm report and choose to either trust or distrust the report based on their experience with the algorithm. Detailed instructions were delivered on the screen following four practice trials. Participants could have their questions answered while instructions were given and during the practice session.

Each trial consisted of: a stimulus (i.e., report on sensor functionality), the participant's response, and feedback to the participants on the correctness of their response. There were two stimuli, 'obstacle detected' and 'clear road', and both had a 50%

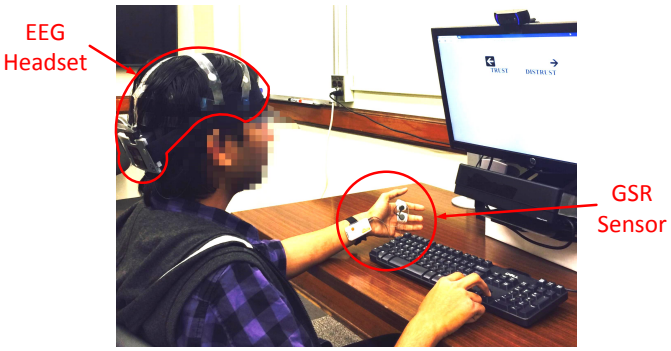


Figure 2.15. Experimental setup with participant wearing EEG Headset and GSR Sensor.

probability of occurrence. Participants had the option to choose 'trust' or 'distrust' in response to the sensor report after which they received the feedback of 'correct' or 'incorrect'. Figure 2.16 shows the sequence of events in a single trial, and Figure 2.17 shows example screenshots of the computer interface.

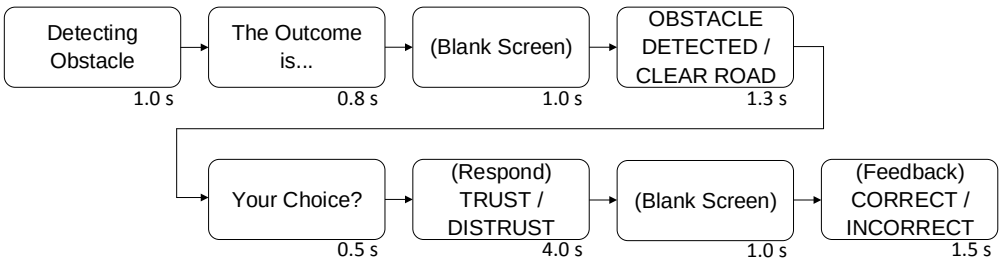
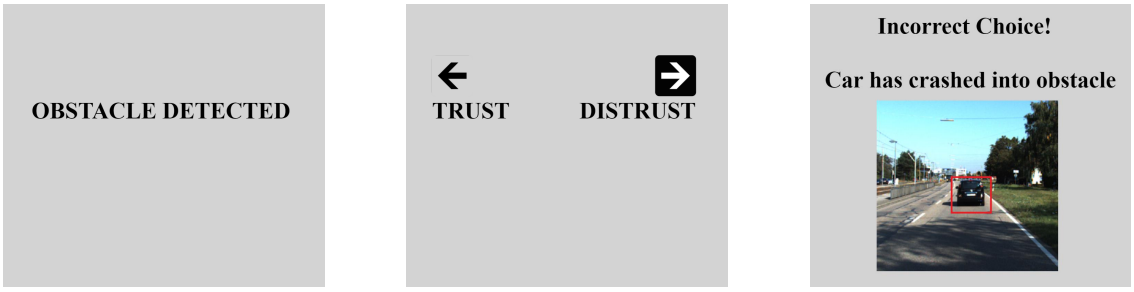


Figure 2.16. Sequence of events in a single trial. The time length marked on the bottom right corner of each event indicates the time interval for which the information appeared on the computer screen.

The independent variable was the participants' experience due to the sensor performance, and the dependent variable was their trust level. The sensor performance was varied to elicit the dynamic response in each participant's trust level. There were two categories of trials: *reliable* and *faulty*. In reliable trials, the sensor accurately



(a) (b) (c)

Figure 2.17. Example screenshots of the interface of the experimental study. The left screenshot (a) shows the stimuli, the middle screenshot (b) shows the response, and the right screenshot (c) shows the feedback. These screens correspond to three of the events shown in Figure 2.16: obstacle detected/clear road, trust/distrust, and correct/incorrect, respectively.

identified the road condition with 100% probability; in faulty trials, there was only a 50% probability that the sensor correctly identified the road condition with sensor faults presented in a randomized order. We implemented the 50% accuracy for faulty trials because pilot studies indicated that it would be perceived as a pure random chance by the participants. This should conceivably result in the lowest possible trust level that a human has in the simulated sensor. The participants received 'correct' as feedback when they indicated trust in reliable trials, but there was a 50% probability that they received 'incorrect' as feedback when they indicated trust in faulty trials.

Each participant completed 100 trials. The trials were divided into three phases, called 'databases' in the study, as shown in Figure 2.18. Participants were randomly assigned to one of two groups for counterbalancing any possible ordering effects. Databases 1 and 2 consisted of either reliable (A) or faulty (B) trials (see details in Figure 2.18). The number of trials in each of these two databases was chosen so that the trust or distrust response of each human subject would approach a steady-state value [41]. Steady-state ensures that the trust level truly reaches the desired state

(i.e., trust for reliable trials and distrust for faulty trials) which is essential for labeling the trials as trust or distrust. On the other hand, the accuracy of the algorithm was switched between reliable and faulty according to a pseudo-random binary sequence (PRBS) in Database 3. This was done in order to excite all possible dynamics of the participant's trust response required for dynamic behavior modeling, which was the subject of related work by the authors [46]. Therefore, only the data from databases 1 and 2 (i.e., the first 40 trials) were analyzed.

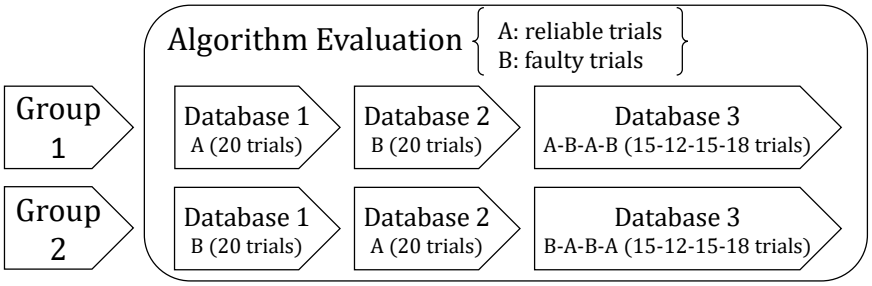


Figure 2.18. Participants were randomly assigned to one of two groups. The ordering of the three experimental sections (databases), composed of reliable and faulty trials, were counterbalanced across Groups 1 and 2.

We collected psychophysiological measurements in order to identify any latent indicators of trust and distrust. In general, latent emotions are those which cannot be easily articulated. Latent distrust may inhibit the interactions between human and intelligent systems despite reported trust behaviors. We hypothesized that the trust level would be high in reliable trials and be low in faulty trials, and we validated this hypothesis using responses collected from 581 online participants (58 were outliers) via Amazon Mechanical Turk [97]. The experiment elicited expected trust responses based on the aggregated data as shown in Figure 2.19 [46]. Therefore, data from reliable trials were labeled as trust, and data from faulty trials were labeled as distrust. The data analysis and feature extraction methodologies will be discussed further in Section 2.2.4.

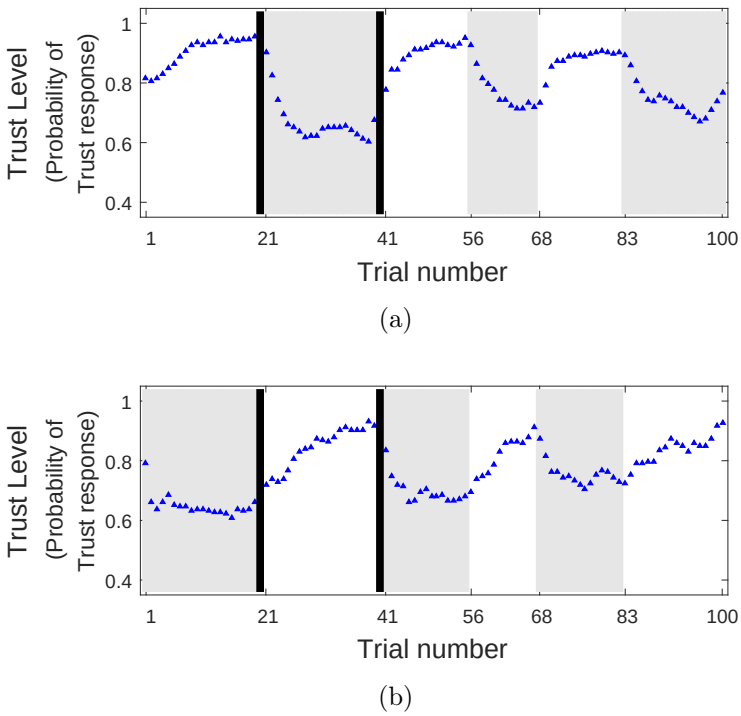


Figure 2.19. The averaged response from online participants collected via Amazon Mechanical Turk. Subfigure (a) corresponds to the 295 participants from group 1 and subfigure (b) corresponds to the 228 participants from group 2. Faulty trials are highlighted in gray. Participants showed a high trust level in reliable trials and a low trust level in faulty trials regardless of the group they were in.

2.2.4 Data Analysis

In this section we discuss the methods used to pre-process the data (collected during the human subject studies) so as to reduce noise and remove contaminated data. We then describe the process of feature extraction applied to the processed data.

Pre-processing

We used the automatic decontaminated signals provided by the B-Alert EEG system for artifact removal. This decontamination process minimizes the effects of electromyography, electrooculography, spikes, saturation, and excursions. Before further processing the data, we manually examined the spectral distribution of EEG data for each participant. We removed the participants having anomalous EEG spectra, possibly due to bad channels or dislocation of EEG electrodes during the study. This process resulted in 45 participants to analyze. Finally, EEG measurements from channel F3 and F4 were excluded from the data analysis due to contamination with eye movement and blinking [109]. For GSR measurements, we used adaptive Gaussian smoothing with a window of size 8 to reduce noise [110].

Feature Extraction

In order to estimate trust in real-time, we require the ability to continuously extract and evaluate key psychophysiological measurements. This could be achieved by continuously considering short segments of signals for calculations. Levy suggests using short epoch lengths for identifying rapid changes in EEG patterns [111]. Therefore, we divided the entire duration of the study into multiple 1-second epochs (periods) with 50% overlap between each consecutive epoch. Assuming that the decisive cognitive activity occurs when the participant sees the stimuli, we only considered the epochs lying completely between each successive stimulus (obstacle detected/clear road) and response (trust/distrust). Consequently, approximately 129 epochs were considered for each participant. We labeled each of these epochs as one of two classes, namely *Distrust* or *Trust*, based on whether the epoch belonged to faulty or reliable trials, respectively. The number of epochs varied depending on the response time of the human subject for each trial.

EEG: Existing studies have shown the importance of both time-domain features and frequency-domain features for successfully classifying cognitive tasks [112]. To utilize the benefits of both, we extracted an exhaustive set of time- and frequency-domain features from EEG.

We extracted six time-domain features from all seven channels (Fz, C3, Cz, C4, P3, POz, and P4) for each epoch of length N. For this study in which EEG signals were sampled at 256 Hz, each 1-second epoch had a length of $N = 256$. Letting $k \in (1, n)$, where n is the total number of epochs and x_k represents the k^{th} epoch of channel ch_x. These features were defined as:

1. mean $\mu_k(ch_x)$, where
$$\mu_k(ch_x) = \frac{1}{N}\sum_{i=1}^{N} x_{ki}, \tag{2.10}$$

2. variance $\sigma_k^2(ch_x)$, where
$$\sigma_k^2(ch_x) = \frac{1}{N-1}\sum_{i=1}^{N} |x_{ki} - \mu_k|^2, \tag{2.11}$$

3. peak-to-peak value $pp_k(ch_x)$, where
$$pp_k(ch_x) = \max_{1\leq i\leq N} x_{ki} - \min_{1\leq i\leq N} x_{ki}, \tag{2.12}$$

4. mean frequency $\bar{f}_k(ch_x)$, defined as the estimate of the mean frequency from the power spectrum of x_k,

5. root mean square value $rms_k(ch_x)$, where
$$rms_k(ch_x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |x_{ki}|^2}, \tag{2.13}$$

and

6. signal energy $E_k(ch_x)$, where

$$E_k(ch_x) = \sum_{i=1}^{N} |x_{ki}|^2 \ . \tag{2.14}$$

Therefore, we extracted 42 (6 features $\times$ 7 channels) time-domain features for each epoch. Moreover, the interaction between the different regions of the brain was also considered by calculating the correlation between pairs of channels for each epoch. The correlation coefficient between two channels (e.g., ch_x and ch_y) of the k^{th} epoch $\rho_k(ch_x, ch_y)$ is defined as

$$\rho_k(ch_x, ch_y) = \frac{cov(x_k, y_k))}{\sqrt{var(x_k)var(y_k)}}, \tag{2.15}$$

where x_k and y_k are the k^{th} epochs of channels ch_x and ch_y respectively. The expressions $cov(.)$ and $var(.)$ are the covariance and variance functions, respectively. Therefore, 21 additional time-domain features were extracted (combinations of 2 out of 7 channels, C_2^7).

Next we extracted features from four frequency bands across all seven channels for each epoch. Classically, EEG brain waves have been categorized into four bands based on frequency, namely, delta (0.5 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 13 Hz), and beta (13 - 30 Hz). However, because of the non-stationary characteristics of EEG signals (i.e., their statistics vary in time), analyzing the variations in frequency components of EEG signal with time (i.e., time-frequency analysis) is more informative than analyzing the frequency content of the entire signal at a time. The Discrete Wavelet Transform (DWT) is an extensively used tool for time-frequency analysis of physiological signals, including EEG [113]. Therefore, we used DWT decomposition to extract the frequency-domain features from the EEG signals.

DWT uses scale-varying basis functions to achieve good time resolution of high frequencies and good frequency resolution for low frequencies. The DWT decomposition consists of successive high pass and low pass filtering of the signal with downsam-

pling by a factor of 2 in each successive level [114]. The high pass filter uses a discrete mother wavelet function, and the low pass filter uses its mirror version. We used the mother wavelet function of the Daubechies wavelet (db5) for frequency decomposition of the EEG signal. The first low pass and high pass filter outputs are called approximation A1 and detailed coefficients D1, respectively. A1 is further decomposed, and the steps are repeated to achieve the desired level of decomposition. Since the highest frequency in our signal was 128 Hz (sampling frequency $f_s = 256$ Hz), each channels' signal was decomposed to the fifth level to achieve the decomposition corresponding to the classical bands as shown in Table 2.4.

Table 2.4.
Wavelet decompositions and their corresponding frequency ranges. The closest classical frequency band for each decomposition is also shown.

Level	Wavelet coefficient	Frequency range	Classical band
3	D3	16 - 32 Hz	Beta
4	D4	8 - 16 Hz	Alpha
5	D5	4 - 8 Hz	Theta
5	A5	0 - 4 Hz	Delta

Three features, namely mean (Equation 2.10), variance (Equation 2.11), and energy (Equation 2.14) were calculated from each of the four decomposed band decomposition coefficients shown in Table 2.4 for each channel's epoch. Therefore, 84 frequency-domain features were extracted (3 features $\times$ 4 bands $\times$ 7 channels).

GSR: GSR is a superposition of the tonic (slow-changing) and the phasic (fast-changing) components of the skin conductance response [115]. We used Continuous Decomposition Analysis from Ledalab to separate the tonic and phasic components of the signal [115]. Since the time-scale of the study and the decision making tasks are, in general, much faster as compared to the tonic component, we only used the phasic component of the GSR. We calculated the *Maximum Phasic Component* and the *Net Phasic Component* for each epoch, thus extracting 2 features from GSR.

2.2.5 Feature Selection

Following the feature extraction described in Section 2.2.4, we next describe the process of feature selection. The selected features were considered to be potential input variables for the trust sensor model, of which the output would be the *probability of trust response*. We define the probability of trust response as the probability of the human trusting the intelligent system at the next time instant. In this section we discuss feature selection algorithms used for selecting optimal feature sets for two variations of our trust sensor model, followed by a discussion of the significance of the features in each of the final feature sets.

Feature Selection Algorithms

The complete feature set consisted of 149 features (42 + 21 + 84 + 2) that were extracted for each epoch for every participant. These features were considered potential variables for predicting the *Trust* or *Distrust* classes. Out of this large feature set, it was necessary to downselect a smaller subset of features as predictors to avoid 'the curse of dimensionality' (also called Hughes phenomenon), which occurs for high-dimensional feature spaces with a limited number of samples. Not doing feature selection leads to a reduction in the predictive power of learning algorithms [112]. Therefore, feature selection was achieved by removing irrelevant and redundant features from the feature set according to feature selection algorithms.

Feature selection algorithms are categorized into two groups: filter methods and wrapper methods. Filter methods depend on general data characteristics such as inter-class distance, results of significance tests, and mutual information, to select the feature subsets without involving any selected prediction model. Since filter methods do not involve any assumptions of a prediction model, they are useful in estimating the relationships between the features. Wrapper methods use the performance (e.g., accuracy) of a selected prediction model to evaluate possible feature subsets. When the performance of a particular type of model is of importance, wrapper methods

result in a better fit for a selected model type; however, they are typically much slower than filter methods [116]. We used a combination of filter and wrapper methods for feature selection to manage the trade-off between training speed and model performance. We used a filter method called *ReliefF* for initially shortlisting features followed by a wrapper method called *Sequential Forward Floating Selection (SFFS)* for the final feature selection as shown in Figure 2.20.

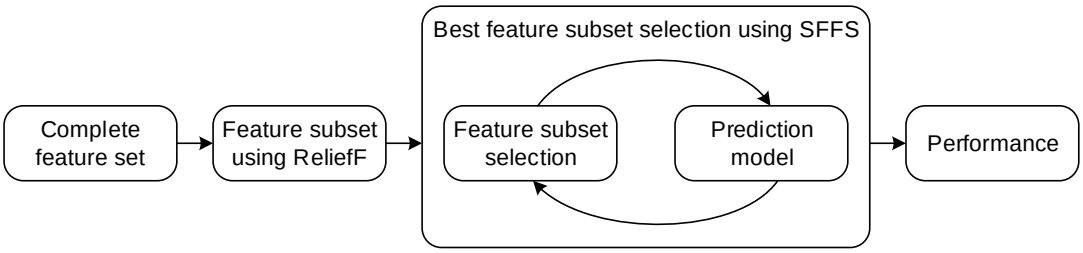


Figure 2.20. A schematic depicting the feature selection approach used for reducing the dimension of the feature set. The ReliefF (filter method) was used for an initial shortlisting of the feature subset followed by SFFS (wrapper method) for the final feature subset selection.

ReliefF: The basic idea of ReliefF is to estimate the quality of the features based on their ability to distinguish between samples that are near each other. Kononenko *et al.* proposed a number of improvements to existing work by Kira and Rendell and developed ReliefF [117, 118]. For a data set with n samples, the algorithm iterates n times for each feature. For our study, there were approximately 129 samples corresponding to each epoch as mentioned in Section 2.2.4. At each iteration for a two-class problem, the algorithm selects one of the samples and finds k nearest hits (same-class sample) and k nearest misses (different-class sample), where k is a parameter to be selected. Kononenko *et al.* suggested that k could be safely set to 10 for most purposes. We used k =10 and calculated the ReliefF weights for all extracted features of each individual participant. The weight of any given feature is penalized

for far-off near-hits and improved for far-off near-misses. Far-off near misses implies well-separated features, and far-off near-hits implies intermixed classes.

Sequential Forward Floating Selection (SFFS): The SFFS is an enhancement of the Sequential Feature Selection algorithm for addressing the 'nesting effect' [119]. The nesting effect means that a selected feature cannot be discarded when the forward method is implemented and the discarded feature cannot be re-selected when the backward method is implemented. In order to avoid this effect, SFFS builds the feature set with the best predictive power by continuously adding a dynamically changing number of features at each step to the existing subset of features. This operation occurs iteratively until no further increase in performance is observed. In this study we defined the performance as the misclassification rate of the Quadratic Discriminant Analysis (QDA) classifier. We have examined that a QDA classifier achieved the highest accuracy for another data set based on the same experimental setup [120], and its output posterior probability is also suitable for interpreting trust. Therefore, we used the QDA classifier and calculated the misclassification rate using 5-fold cross validation [121]. This validation technique randomly divides the data into five sets and predicts each set using a model trained for the remaining four sets.

Feature Selection for the Trust Sensor Model

The differences between humans could introduce differences in their trust behavior. This leads to two approaches for selecting features for sensing trust level: 1) to select a common set of features for a general population, which results in a *general trust sensor model*; and 2) to select a different set of features for each individual, which results in *customized trust sensor model* for each individual.

Feature Selection for the General Trust Sensor Model: A general trust sensor model is desirable so that it can be used to reflect trust behavior in a general adult population. This model correlates significant psychophysiological features with

human trust in intelligent systems based on data obtained from a broad range of adult human subjects. Since a general trust sensor model requires a common list of features for all participants, we randomly divided the participants into two groups: the training-sample participants (33 out of 45 participants), which were used to identify the common list of features, and the validation-sample participants (12 out of 45 participants), which were used to validate the selected list of features. We calculated the median of the ReliefF weights across the training-sample participants for all features. The median was used instead of mean to avoid outliers [122]. Finally, we shortlisted features with the top 60 median weights and used SFFS for selecting the final set of features. For each training-sample participant's data, a separate classifier was trained and the average value of the misclassification rate for all training-sample participants was used as the predictive power for feature subsets for SFFS. We obtained a feature set with 12 features consisting of both time- and frequency-domain features of EEG along with net phasic components of GSR. Table 2.5 shows the final list of selected features for the general trust sensor model using training-sample participants.

Table 2.5.
Features to be used as input variables for the general trust sensor model

	Feature	Measurement	Domain
1	Mean Frequency - Fz	EEG	Time
2	Mean Frequency - C3	EEG	Time
3	Mean Frequency - C4	EEG	Time
4	Peak-to-peak - C3	EEG	Time
5	Energy of Theta Band - P3	EEG	Frequency
6	Variance of Alpha Band - P4	EEG	Frequency
7	Energy of Beta Band - C4	EEG	Frequency
8	Energy of Beta Band - P3	EEG	Frequency
9	Mean of Beta Band - C3	EEG	Frequency
10	Correlation - C3 & C4	EEG	Time
11	Correlation - Cz & C4	EEG	Time
12	Net Phasic Component	GSR	Time

Feature Selection for the Customized Trust Sensor Model: We followed a similar approach to that used for feature selection in Section 2.2.5, but the list of features was selected individually for each of the 45 participants. We used ReliefF weights and shortlisted a separate set of features for each participant consisting of the top 60 weights. Then, for each participant, SFFS was used with the misclassification rate as determined by the quadratic discriminant classifier to select a final set of features from the shortlisted feature set. We obtained a relatively smaller feature set for each individual participant, with an average of 4.33 features in each participant's feature set, as compared to 12 features when all of the participants' data was aggregated into a single data set. Table 2.6 shows each of the features that are significant for at least four of the participants. We observed that there is great diversity in the significant features for each individual which supports the usage of a customized trust sensor model. However, it is important to note that even within this diversity, more than half of the most common features (e.g., mean frequency at C4) are also significant for the general trust sensor model.

Table 2.6.
The most common features that are significant for at least four participants. Features marked with an asterisk (*) are also significant for the general trust sensor model.

	Feature	Measurement	Domain
1	Mean Frequency - POz	EEG	Time
2	Mean Frequency - C4*	EEG	Time
3	Mean Frequency - P3	EEG	Time
4	Mean Frequency - Fz*	EEG	Time
5	Mean Frequency - C3*	EEG	Time
6	Peak-to-peak - C3*	EEG	Time
7	Variance of Beta Band - P3	EEG	Frequency
8	Mean of Beta Band - P3	EEG	Frequency
9	Correlation - Cz & C4*	EEG	Time
10	Net Phasic Component*	GSR	Time
11	Maximum Value of Phasic Activity	GSR	Time

Discussion on Significant Features in Trust Sensing

Several time-domain EEG features were found to be significant, especially the mean frequency of the EEG power distribution and the correlations between the signals from the central regions of the brain (C3, C4, Cz). Time-domain EEG features have been discovered to be significant in brain activities [112]. Moreover, our observation that activities at sites C3 and C4 play an important role in trust behaviors is supported by existing studies that have suggested that central regions of the brain are related to processes associated with problem complexity [123], anxiety in a sustained attention task [124], and mental workload [125].

Among the frequency domain EEG features, the measurements from the left parietal lobe, particularly in a high frequency range (i.e., the beta band), responded most strongly to the discrepancy between reliable and faulty stimuli. This is consistent with the finding that cognitive task demands have a significant interaction with hemisphere in the beta band for parietal areas [126]. The beta band is also an important feature that has been shown to be related to emotional states in the literature [127] and may represent the emotional component of human trust.

Finally, the results also showed that the phasic component of GSR was a significant predictor of trust levels for the general trust sensor model as well as for several customized trust sensor models. This aligns with the existing literature that shows that the GSR features could significantly improve the classification accuracy for mental workload detection [128] and could index difficulty levels of decision making [129]. The importance of phasic GSR to trust sensing was also supported by Khawaji's study in which the average of peak GSR values was affected by interpersonal trust [43].

2.2.6 Model Training and Validation

The selected features discussed in Section 2.2.5 were considered as input variables for each of the trust sensor models; the output variables were the categorical trust level, namely the classes 'Trust' and 'Distrust'. In this section we introduce the

training procedure of a quadratic discriminant classifier that was used to predict the categorical trust class using the psychophysiological features. We then present and discuss the results of the model validation.

Classifier Training

The quadratic discriminant classifier was implemented using the Statistics and Machine Learning Toolbox in MATLAB R2016a (The MathWorks, Inc., USA). The low training and prediction time of quadratic discriminant classifiers is advantageous for real-time implementation of the classifier [130]. Moreover, the posterior probability calculated by the classifier for the class 'Trust' was used as the probability of trust response, thus resulting in a continuous output. The continuous output of probability of trust response would be particularly beneficial for implementation of a feedback control algorithm for managing human trust level in an intelligent system. In order to avoid large and sudden fluctuations in the trust level, the continuous output was smoothed using a median filter with a window of size 15. The general trust sensor model and customized trust sensor models were developed with the same training procedure but with different feature sets (i.e., input variables). The former was based on the common feature set, and the latter was based on customized feature sets, as described in Sections 2.2.5 and 2.2.5.

Model Validation Techniques

We used 5-fold cross-validation to evaluate the performance of classifiers. The data, consisting of approximately 129 samples for each participant, was randomly divided into 5 sets. Each set was predicted using a model trained from the other four datasets. We used these predictions to evaluate the accuracy of the binary classification. Accuracy is defined as the proportion of correct predictions among the total number of samples and is given as

$$\text{accuracy} = \frac{\text{Correct Predictions}}{\text{Total population}} \quad . \tag{2.16}$$

Moreover, prediction performance of a classifier may be better evaluated by examining the confusion matrix shown in Figure 2.21. We calculated two statistical measures called sensitivity (true positive ratio) and specificity (true negative ratio) that are defined as follows.

1. Sensitivity: the proportion of actual trust (positives) that are correctly predicted as such, where

$$\text{sensitivity} = \frac{\text{True positives}}{\text{True positives} + \text{False negatives}} \quad . \tag{2.17}$$

2. Specificity: the proportion of actual distrust (negatives) that were correctly predicted as such, where

$$\text{specificity} = \frac{\text{True negatives}}{\text{True negatives} + \text{False positives}} \quad . \tag{2.18}$$

In order to examine the robustness of the classifier to the variation in training data, we performed 10,000 iterations with a different random division of the five sets in each iteration and calculated the performance measures for each iteration. Table 2.7 and Table 2.8 show the mean, maximum (Max), minimum (Min), and standard deviation (SD) values for each of the performance measures for the *general trust sensor model*. This is shown for both training-sample participants (Table IV) and validation-sample participants (Table V) along with the 95% confidence interval (CI) obtained using the iterations. Table 2.9 shows the performance statistics of the *customized trust sensor model* for all participants. The confidence intervals obtained for both models were very narrow, *indicating that models were robust to the selection of training data.*

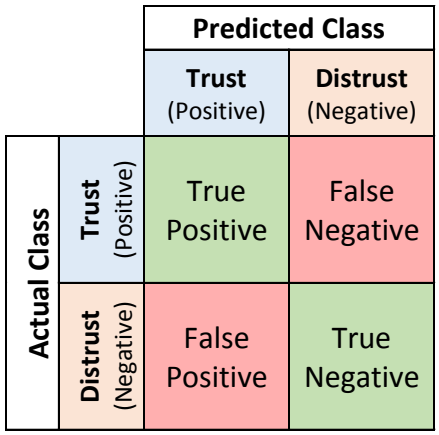


Figure 2.21. The actual class and the predicted class form a 2×2 confusion matrix. The outcomes are defined as true or false positive/negative.

Table 2.7.
The accuracy, sensitivity, and specificity (%) of the *general* trust sensor model for training-sample participants with a 95% confidence interval

	Accuracy	Sensitivity	Specificity
Mean	70.52 ± 0.007	64.17 ± 0.010	75.49 ± 0.009
Max	93.72 ± 0.013	96.75 ± 0.020	96.38 ± 0.015
Min	54.67 ± 0.042	31.18 ± 0.040	44.92 ± 0.039
SD	11.29 ± 0.006	18.96 ± 0.009	14.35 ± 0.008

Table 2.8.
The accuracy, sensitivity, and specificity (%) of the *general* trust sensor model for validation-sample participants with a 95% confidence interval

	Accuracy	Sensitivity	Specificity
Mean	73.13 ± 0.010	65.35 ± 0.015	79.49 ± 0.013
Max	99.89 ± 0.006	99.92 ± 0.006	99.85 ± 0.011
Min	59.29 ± 0.035	34.35 ± 0.081	57.04 ± 0.050
SD	10.91 ± 0.007	17.03 ± 0.016	12.26 ± 0.015

Table 2.9.
The accuracy, sensitivity, and specificity (%) of the *customized* trust sensor model for all participants with a 95% confidence interval

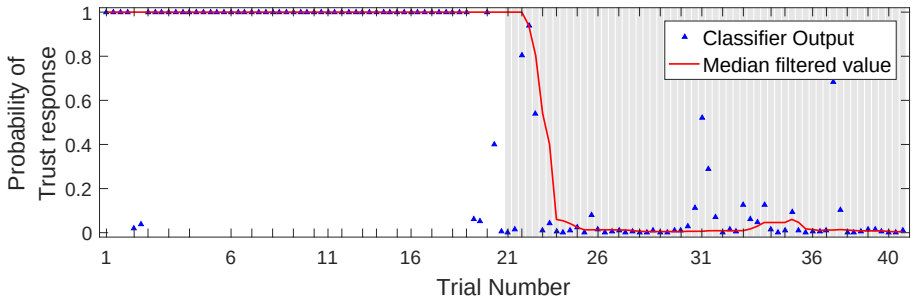
	Accuracy	Sensitivity	Specificity
Mean	78.55 ± 0.005	72.83 ± 0.007	82.56 ± 0.007
Max	100.00 ± 0.000	100.00 ± 0.000	100.00 ± 0.000
Min	61.59 ± 0.041	34.77 ± 0.044	45.89 ± 0.040
SD	9.69 ± 0.005	17.02 ± 0.008	11.18 ± 0.007

Discussion on Performance of Classification Models

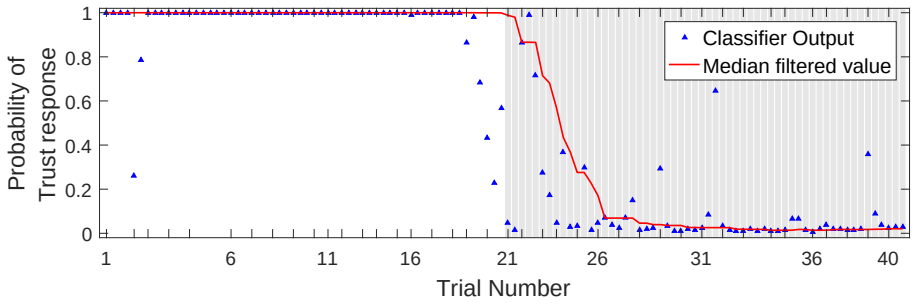
The mean accuracy was 70.52±0.007% for training-sample participants. Similarly, the mean accuracy for the *validation-sample* participants was 73.13±0.010%. The fact that the performance of the general trust model was consistent for both training-sample and validation-sample participants suggests that the identified list of features could estimate trust for a broad population of individuals. Moreover, the mean accuracy was 78.58±0.0005% for the customized trust sensor models for all participants. Recall that the customized trust senor models were based on a customized feature set for each participant. There were 12 significant features to predict trust for the general trust sensor models, while less than 5 features were needed for the customized trust sensor models. These findings support the hypothesis that a customized trust sensor model could enhance the prediction accuracy with a smaller feature set. For some individual participants, the mean accuracy increased to 100%.

Figures 2.22 and 2.23 are examples of good predictions for participants in groups 1 and 2, respectively. The customized trust sensor models performed better for both participants, specifically at the transition state at the beginning of database 2. Figure 2.22(b) shows an example of a transition state at the beginning of database 2; it took five trials for this participant to establish a new trust level. The classification accuracy was low for some participants as shown in Figure 2.24. The classifier had difficulty correctly predicting trust (database 1), which may imply that this partic-

ular participant was not able to conclude whether or not to trust the sensor report, even in reliable trials. Another potential reason could be that trust variations of this participant did not result in significant changes in their physiological signals. Nevertheless, the customized trust sensor model still showed a higher accuracy than the general trust sensor model.



(a)



(b)

Figure 2.22. Classifier predictions for participant 44 in group 1. The top figure (a) shows the general trust sensor model predictions with an accuracy of 90.52%. The bottom figure (b) shows the customized trust sensor model predictions with an accuracy of 93.97%. Faulty trials are highlighted in gray. Trust sensor models had a good accuracy for this participant. The classifier output of posterior probability was smoothed using a median filter with window of size 15.

The general trust sensor model resulted in mean *specificity* of 75.49±0.009% and 79.49±0.013% for training-sample and validation-sample participants, respectively. The customized trust sensor model resulted in 82.56±0.007% for all participants. This indicates that the models are capable of correctly predicting distrust in hu-

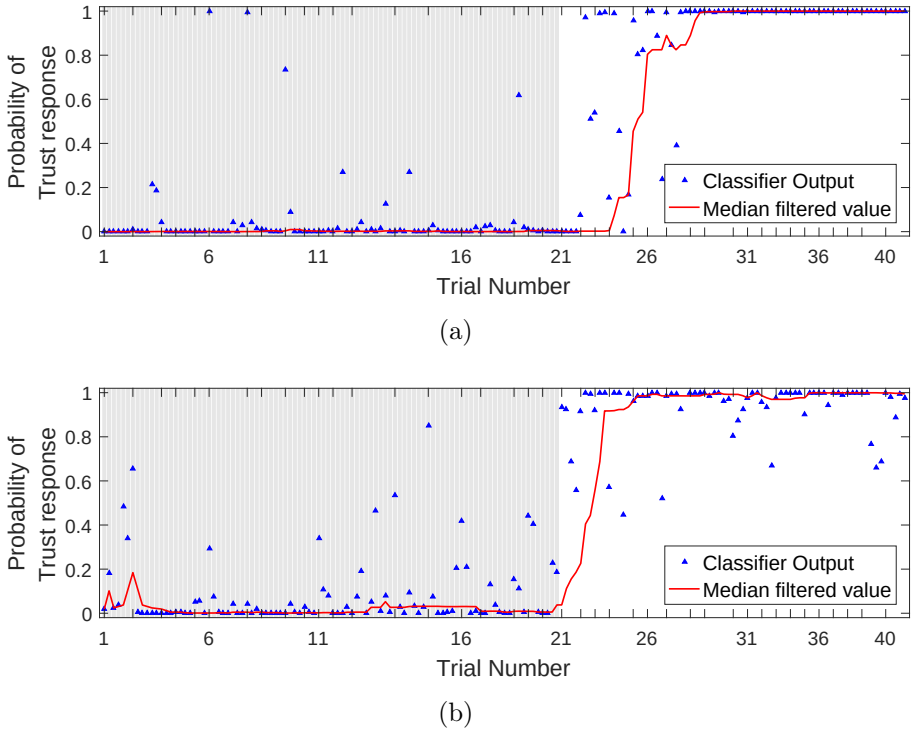


Figure 2.23. Classifier predictions for participant 10 in group 2. The top figure (a) shows the general trust sensor model predictions with an accuracy of 91.12%. The bottom figure (b) shows the customized trust sensor model predictions with an accuracy of 96.45%. Faulty trials are highlighted in gray. Trust sensor models had good accuracy for this participant. The classifier output of posterior probability was smoothed using a median filter with window of size 15.

mans. The models are less likely to predict a distrust response as trust (i.e., less false positives). The mean *sensitivity* was 64.17±0.010% and 65.35±0.015% for the general trust sensor model for training-sample and validation-sample participants, respectively. The customized trust sensor model resulted in 72.83±0.007% for all participants. Low sensitivity (more false negatives) occurs when the model often predicts trust as distrust. In the context of using this trust sensor model to design an intelligent system that could be responsive to a human's trust level, low sensitivity would arguably not have an adverse effect since the goal of the system would be to enhance trust.

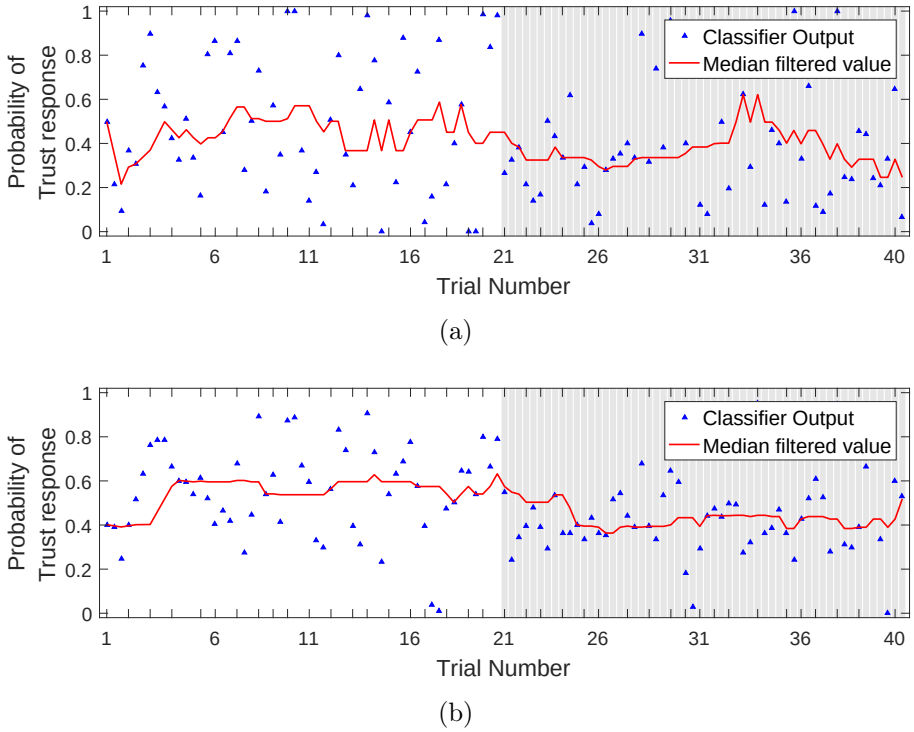


Figure 2.24. Classifier predictions for participant 8 in group 1. The top figure (a) shows the general trust sensor model predictions with an accuracy of 61.26%. The bottom figure (b) shows the customized trust sensor model predictions with an accuracy of 72.07%. Faulty trials are highlighted in gray. Trust sensor models did not have good accuracy for this participant. The classifier output of posterior probability was smoothed using a median filter with window of size 15.

There is a fundamental trade-off that exists between the general and customized models in terms of the time spent on model training and model performance as shown in Table 2.10. The results show that the selected feature set (Table 2.5) for the general trust sensor models is applicable for a general adult population with a 71.22% mean accuracy (i.e., the mean accuracy calculated across all participants). Furthermore, by applying this common feature set, feature selection is not required while implementing the general model. This would reduce the model training time and potentially make the model adaptable to various scenarios. However, the common feature set for a general population is larger than feature sets optimized for each individual because it

attempts to accommodate an aggregated group of individuals. Therefore, in scenarios where the speed of the online prediction process is the priority, the customized trust sensor model, with a smaller feature set, would be preferred. The customized trust sensor model also enhances the prediction accuracy. Nonetheless, it is worth noting that implementing the customized trust sensor model would still require extraction of a larger set of features initially for training followed by a smaller feature set extraction for real-time implementation. This would increase the time required for training the model as an additional feature selection step would need to be performed.

While we focused on situational and learned trust, dispositional trust factors, such as demographics, may have partially contributed to the observed lower accuracy of the general trust sensor model due to individual differences in trust response behavior [46, 131]. Incorporating these additional factors and other psychophysiological signals may increase the trust estimation accuracy of the trust sensor model, as the features included in the present model inherently represent only a subset of many non-verbal signals that correlate to trust level.

In summary, the proposed trust sensor model could be used to enable intelligent systems to estimate human trust and in turn respond to, and collaborate with, humans in such a way that leads to successful and synergistic collaborations. Potential human-machine/robot collaboration contexts include robotic nurses that assist patients, aircrafts that exchange control authority with human operators, and numerous others [1].

The results presented in this section show that psychophysiological measurements can be used to estimate human trust in intelligent systems in real-time. We proposed two approaches for developing classifier-based empirical trust sensor models that estimate human trust level using psychophysiological measurements. These models used human subject data collected from 45 participants. The first approach was to consider a common set of psychophysiological features as the input variables for any human and train a classifier-based model using this feature set, resulting in a general trust sensor model with a mean accuracy of 71.22%. The second approach was to consider

Table 2.10.
Comparison of General Trust Sensor Model and Customized Trust Sensor Model for implementation

Model Characteristics	General Trust Sensor Model	Customized Trust Sensor Model
Required training time	Less	More
Size of final feature set	12	4.33 (Average)
Prediction Time	More	Less
Mean Prediction Accuracy	71.22%	78.55%

a customized feature set for each individual and train a classifier-based model using that feature set; this resulted in a mean accuracy of 78.55%. The primary trade-off between these two approaches was shown to be training time and performance (based on mean accuracy) of the classifier-based model. That is to say, while it is expected that using a feature set customized to a particular individual will outperform a model based upon the general feature set, the time needed for training such a model may be prohibitive in certain applications. Moreover, although the criteria used for feature selection and classifier training in this study was mean accuracy, a different criterion could be chosen to adapt to various applications. In the next section, we present an adaptive classification framework that combines psychophysiological measurements and human behavioral dynamics to estimate human trust in real time.

2.3 Combining Behavioral and Psychophysiological Measurements

The contents of this section were previously published by Akash, Reid, and Jain in *the Proceedings of the 2018 American Control Conference (ACC)* [132] and are reported here with minor modifications.

2.3.1 Introduction

In the last section, we discussed the use of classification algorithms to estimate human trust using psychophysiological measurements. However, in the application of most classification algorithms, it is assumed that data samples are independent, identically distributed, and are characterized by a stationary distribution. Numerous classification algorithms have been developed for data that satisfy these assumptions (see [133] for a review). However, many real-world problems are characterized by data with temporal variations and a non-stationary distribution. One example is the use of human behavioral responses and psychophysiological data for prediction of human behavior, in particular, human trust. In this section, we present an adaptive probabilistic classification algorithm which incorporates the temporal dynamics of the human trust behavior with EEG measurements to estimate trust.

Human behavior and emotion estimation is becoming an important segment in the fields of modern human-machine interaction, brain-computer interface (BCI) design, and medical care [134], among others. Human behavior inference for decision making is critical for building synergistic relationships between humans and autonomous systems. Researchers have attempted to predict human behavior using dynamic models that rely on the behavioral responses or self-reported behavior of humans [33, 54]. An alternative is the use of psychophysiological signals like the electroencephalogram (EEG) that represents the electrical activity of the brain. In order to infer human behavior from psychophysiological signals, different brain activity patterns must be identified. A common approach for this identification is the use of classification algorithms [112]. However, most of the EEG-based classification algorithms in literature are based on static classifiers that do not account for the dynamic characteristics of human behavior [112]. Therefore, our goal is to use both behavioral responses and psychophysiological measurements to create a more accurate and robust classification algorithm that considers the dynamics of human behavior.

Most existing classification algorithms do not consider the temporal dynamics of the process under consideration. For classification of dynamic processes such as human behavior, inclusion of the temporal dynamics will improve prediction accuracy. However, dynamic classification algorithms (e.g., hidden Markov models) are typically computationally expensive to train adaptively, and therefore, cannot be used for data with non-stationary characteristics [135–137].

In this section, we present an adaptive probabilistic classification algorithm which incorporates the temporal dynamics of the underlying process under consideration. We use a generative model with the prior probability modeled using a Markov decision process and the conditional probability modeled using an existing adaptive quadratic discriminant analysis classifier. We implement the proposed algorithm for classification of human trust in automation using psychophysiological measurements along with human behavioral responses. Finally, we cross-validate the classifier and show the improvement in its performance as compared to the adaptive classification algorithm alone.

This section is organized as follows. Section 2.3.2 provides background on classification algorithms using EEG. The proposed classification model framework is described in Section 2.3.3. The implementation of the proposed model for predicting human trust is presented in Section 2.3.4. Results and discussions are presented in Section 2.3.5.

2.3.2 Background

There are several classification algorithms which are used in BCI applications and human behavior predictions. These include a variety of algorithms, including linear classifiers (e.g. linear discriminant analysis, support vector machines), nonlinear Bayesian classifiers, artificial neural networks, and k-nearest neighbors [112]. These classifiers can be categorized using two taxonomies: Generative vs. Discriminative and Static vs. Dynamic.

Generative classifiers, e.g., Bayes quadratic discriminant analysis (QDA), learn the distribution of each class and compute the likelihood of each class for classification. Discriminative classifiers, e.g., support vector machines (SVM), only learn the explicit decision boundaries between the classes, which are then used for classification [138]. Since the EEG signals have non-stationary distributions, data collected on-line may be characterized by different underlying distributions than the training data. Therefore, for an adaptive implementation, it is preferable to identify the changes in the underlying distribution and update a generative model accordingly than to update the decision boundary in a discriminative classifier. Furthermore, generative models are typically specified as probabilistic models; this enables a richer description between features and classes than can be achieved using discriminative models by providing a distribution model of how the data are actually generated.

Static classifiers, e.g., SVM, do not account for temporal information during classification as they classify a single feature vector. In contrast, dynamic classifiers, e.g., hidden Markov models (HMM), account for temporal dynamics by classifying a sequence of feature vectors. HMMs have been used for classification of temporal sequences of EEG features as described in [135–137]. While these studies showed that they were promising classifiers for BCI systems, the Viterbi algorithm used for training HMM is both computationally expensive and memory intensive [139]. Therefore, HMM is undesirable for use as an adaptive algorithm. Instead, to design an adaptive probabilistic classifier, we will use a generative model, namely, the Bayesian quadratic discriminant analysis (QDA) classifier. To include temporal dynamics in the classification, we propose to supplement the QDA classifier with a dynamic behavioral model using Markov decision process.

2.3.3 Probabilistic Classification Algorithm

Probabilistic classifiers predict a probability distribution over the classes, instead of just predicting the most likely class. For predicting the probability of a class

label $\mathcal{C}_k$ using the feature vector $\boldsymbol{x}$, we use training data to learn a model for the posterior class probability $P(\mathcal{C}_k|\boldsymbol{x})$. A subsequent decision state uses these posterior class probabilities to assign class labels. *Generative models* initially determine the class-conditional probabilities $P(\boldsymbol{x}|\mathcal{C}_k)$ for each class $\mathcal{C}_k$ and also presume the prior class probabilities $P(\mathcal{C}_k)$. Then, they use Bayes' theorem,

$$P(\mathcal{C}_k|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\mathcal{C}_k)P(\mathcal{C}_k)}{P(\boldsymbol{x})} \tag{2.19}$$

to estimate the posterior class probabilities $P(\mathcal{C}_k|\boldsymbol{x})$. The denominator $P(\boldsymbol{x})$ is a normalization constant.

We consider generative models in this work and incorporate dynamic characteristics using the prior class probabilities based on Markov decision process as discussed in Section 2.3.3. In Section 2.3.3, we provide the mathematical foundations for the QDA classifier as well as an adaptive implementation of it based on [140].

Adaptive Quadratic Discriminant Analysis Classifier

A Quadratic Discriminant Analysis (QDA) classifier uses a generative approach for classification. The posterior probability that a point $\boldsymbol{x}$ belongs to class $\mathcal{C}_k$ is calculated using (2.19) as the product of the prior probability ($P(\mathcal{C}_k)$) and the multivariate normal density ($P(\boldsymbol{x}|\mathcal{C}_k)$) [141]. The density function of the multivariate normal distribution with mean $\boldsymbol{\mu_k}$ and covariance $\boldsymbol{\Sigma_k}$ at a point $\boldsymbol{x}$ is

$$P(\boldsymbol{x}|\mathcal{C}_k) = \frac{1}{\sqrt{2\pi|\boldsymbol{\Sigma_k}|}} \exp\Big(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})^T\boldsymbol{\Sigma_k}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_k})\Big) \ , \tag{2.20}$$

where $|\boldsymbol{\Sigma_k}|$ is the determinant of $\boldsymbol{\Sigma_k}$ [141]. The Quadratic Discriminant Analysis (QDA) classifies $\boldsymbol{x}$ to a class $\mathcal{C}_k$ so as to maximize a posteriori probability of the class, i.e.,

$$\hat{\mathcal{C}_k} = \underset{i=1,\ldots,K}{\operatorname{argmax}} \hat{P}(\mathcal{C}_i|\boldsymbol{x}) \ . \tag{2.21}$$

Therefore, to train a QDA classifier, we need to estimate the means ($\boldsymbol{\mu_k}$) and covariance matrices ($\boldsymbol{\Sigma_k}$) for each class label. This estimation is given by the Maximum Likelihood Estimate (MLE) as $\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$, and $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T - \hat{\boldsymbol{\mu}}^2$. Moreover, the prior probabilities for each class, $P(\mathcal{C}_k)$, are estimated using the sample frequency of each class in the training data. The parameters are typically estimated using a training dataset offline and then used for prediction. However, an adaptive implementation of the QDA classifier developed by Anagnostopoulos et al. [140] uses online learning with forgetting factor λ as shown in (2.22).

$$\hat{\boldsymbol{\mu}}_t = \Big(1 - \frac{1}{t}\Big)\hat{\boldsymbol{\mu}}_{t-1} + \frac{1}{t}\boldsymbol{x}_t, \hat{\boldsymbol{\mu}}_0 = 0 \tag{2.22a}$$

$$\hat{\boldsymbol{\Pi}}_t = \Big(1 - \frac{1}{t}\Big)\hat{\boldsymbol{\Pi}}_{t-1} + \frac{1}{t}\boldsymbol{x}_t\boldsymbol{x}_t^T, \hat{\boldsymbol{\Pi}}_0 = 0 \tag{2.22b}$$

$$\hat{\boldsymbol{\Sigma}}_t = \hat{\boldsymbol{\Pi}}_t - \hat{\boldsymbol{\mu}}_t\hat{\boldsymbol{\mu}}_t^T \tag{2.22c}$$

$$n_t = \lambda_{t-1}n_{t-1} + 1 \tag{2.22d}$$

Here, $\bullet_t$ refers to the t^{th} discrete time value of the variable $\bullet$. The prior probabilities can be calculated as

$$\big(P(\mathcal{C}_k)\big)_t = \Big(1 - \frac{1}{n_t}\Big)\big(P(\mathcal{C}_k)\big)_{t-1} + \frac{1}{n_t}\mathbb{I}((\mathcal{C}_k)_t = \mathcal{C}_k), \tag{2.23}$$

where $\mathbb{I}(x = k)$ is the indicator function that is equal to 1 when the value of x is equal to that of k; else it is 0. A complete derivation can be found in [140].

Dynamic probabilistic model for prior probability

Apart from model adaptation, the adaptive QDA classifier is static in nature; that is, the classifier only considers the present data without considering the *dynamics* of the data. Though past data could be used as a part of $\boldsymbol{x}$, it would significantly increase the dimension of parameters to be estimated. Instead, we propose a dynamic probabilistic model to estimate the prior probability $P(\mathcal{C}_k)$ that would supplement

the estimation of posterior probability $P(\mathcal{C}_k|\boldsymbol{x})$ using (2.19). The input to this model could include variables from $\boldsymbol{x}$ and/or other variables that were not used for the classifier. The modeling frameworks for this dynamic probabilistic model can include state space models (SSM), Markov decision processes (MDP), or HMMs. Here we will consider the use of MDP for modeling the prior probability for classification.

A MDP is a 5-tuple $(\mathcal{S}, \mathcal{A}, T, \mathcal{R}, \gamma)$, with a finite set of states $\mathcal{S}$, a finite set of actions $\mathcal{A}$, state transition probability function $T(s'|s,a) = P[S_{t+1} = s'|S_t = s, A_t = a]$, reward function $\mathcal{R}$, and discount factor $\gamma \in [0,1]$. MDPs are typically used for reinforcement learning to identify the best policy that maximizes the reward. Policy identification is outside the scope of this work. Therefore, for our application of probabilistic dynamic modeling, the reward function $\mathcal{R}$ and the reward discount factor γ will not be considered.

If $T(s'|s,a)$ is not known, it can be empirically estimated, based upon data consisting of actions and corresponding state transitions, using the MLE given as

$$\hat{T}(i,j,k) = \frac{N_{ijk}}{\sum_j N_{ijk}} \tag{2.24}$$
$$N_{ijk} = \sum_{t=1}^{n} \mathbb{I}(s_t = i)\mathbb{I}(s_{t+1} = j)\mathbb{I}(a_t = k) \ ,$$

where $\mathbb{I}(s_t = i)$ is the indicator function which is equal to 1 when the state s at time t is i, else it is 0. The other two indicator functions are similarly defined. Once the state transition probability function $T(s'|s,a)$ is known, the probability for the next state s' is based on the present state s and action a as $T(S_t = s, S_{t+1} = s', A_t = a)$. Further, the n step ahead transition matrix T_n can be calculated given the series of actions $a_t, a_{t+1}, ..., a_{t+i}, ..., a_{t+n-1}$, as

$$T_n = \prod_{i=0}^{n-1} T(:,:,a_{t+i}) \ , \tag{2.25}$$

and thereafter, the n-step ahead probabilities of states p_n can be calculated as $p_n = p_0 T_n$, where p_0 are the initial probabilities of states. These probabilities p_n will be used

as the prior probability $P(\mathcal{C}_k)$ in (2.19) with each state s of the MDP corresponding to the labels $\mathcal{C}_k$ in the QDA classifier.

2.3.4 Classification of Human Trust in HMI

In this section, we describe the classification of human trust behavior using psychophysiological measurements of participants, specifically EEG, along with their behavioral responses. We used behavioral responses to model the prior probability $P(\mathcal{C}_k)$ as described in Section 2.3.3. The features extracted from the psychophysiological measurements were then used as the input $\boldsymbol{x}$ for the adaptive QDA model described in Section 2.3.3. The framework for our adaptive classification model for human trust is shown in Fig. 2.25.

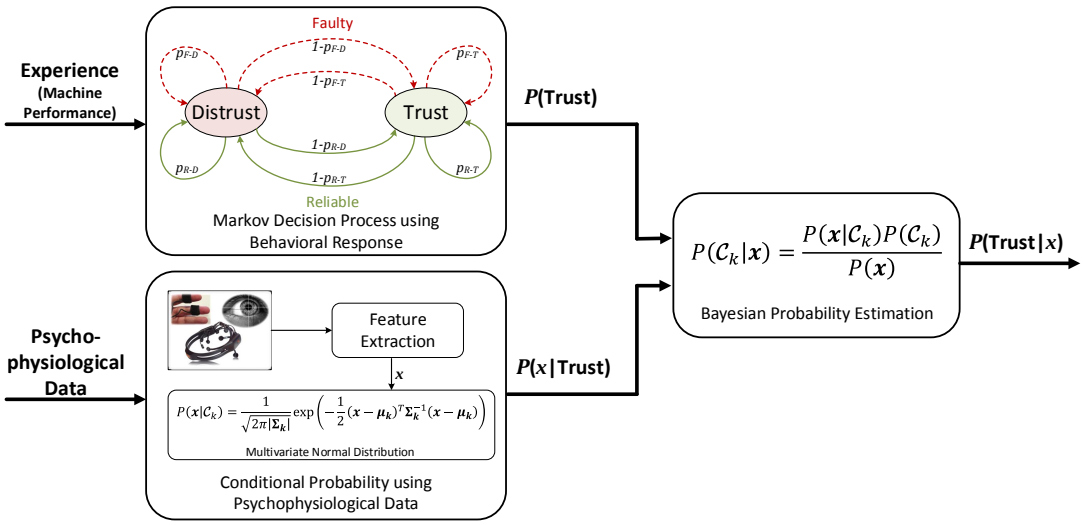


Figure 2.25. A framework for adaptive probabilistic classification of human dynamic trust behavior. A Markov decision process model is used for estimating prior probability using the behavioral responses of participants. Psychophysiological measurements from the participants are used for estimating the conditional probability for each trust state.

Methods and Procedures

In Sections 2.1 and 2.2, we developed an experiment to elicit human trust dynamics in a simulated autonomous system. The participants interacted with a computer interface in which they were told that they would be driving a car equipped with an image-based *obstacle detection sensor*. The sensor would detect obstacles on the road in front of the car, and the participant would need to evaluate the algorithm reports and choose to either trust or distrust the report based on their experience with the algorithm. The study used a within-subjects design with respect to trust wherein both behavioral and psychophysiological data were collected. We used the data to estimate and validate the classification model for each participant. A detailed description of the study design and methods is presented in Sections 2.1 and 2.2.

Five hundred eighty-one participants (340 males, 235 females, and 6 unknown) recruited using Amazon Mechanical Turk [97], participated in our study online. The compensation was \$0.50 for their participation, and each participant electronically provided their consent. The Institutional Review Board at Purdue University approved the study. These data only consisted of the behavioral responses and were used to estimate the MDP model parameters.

Forty-eight adults between 18 and 46 years of age (mean: 25.0 years old, standard deviation: 6.9 years) from West Lafayette, Indiana (USA) were recruited using fliers and email lists and participated in an in-lab study. All participants were compensated at a rate of \$15/hr. The group of participants were diverse with respect to their age, professional field, and cultural background (i.e., nationality). Psychophysiological data along with behavioral data were collected from these participants and used for modeling and validation of the proposed trust classification algorithm. We removed data for three participants that had anomalous EEG spectra, possibly due to bad channels or dislocation of EEG electrodes during the study, resulting in 45 participants to analyze.

Trust behavior modeling using MDP

At each trial, each participant was presented with a stimuli (obstacle detected or clear road) to which they had to respond 'trust' or 'distrust' based on their previous experience (reliable or faulty trial) and from the feedback they received about the sensor after they responded. For this experiment, we define human trust behavior as the process we will model using an MDP as described below:

- The trust decision of the humans is the finite set of states, i.e., $\mathcal{S} : \{\text{Distrust}, \text{Trust}\}$
- The decision process of human trust is influenced by the actions of the machine that lead to the machine performance (experience) as the finite set of actions, i.e., $\mathcal{A} : \{\text{Reliable}, \text{Faulty}\}$
- The experience from trial t acts as an action for the new process state at $t+1$. Therefore, the human state s of trust at t moves to a new state s' at $t+1$ due to the action (i.e., machine performance or experience) at t.
- The state transition probability function $T(s'|s,a)$ can be represented as a $2 \times 2 \times 2$ matrix, such that $T(i,j,k)$ represents the transition probability from i^{th} state to j^{th} state given the action k. Therefore, each of $P(:,:,k)$ represents the state transition matrix for the k^{th} action.

We estimated the transition probability function as well as the initial state probabilities using the behavioral data collected from Amazon Mechanical Turk. We used an aggregated data of 581 participants for the estimation, and therefore assumed that a single transition probability function is representative of general human trust behavior. The estimated probability matrices are given as

$$T(s, s', a = \text{Faulty}) = \begin{bmatrix} 0.5343 & 0.4857 \\ 0.3131 & 0.6869 \end{bmatrix},$$
$$T(s, s', a = \text{Reliable}) = \begin{bmatrix} 0.3177 & 0.6823 \\ 0.1191 & 0.8809 \end{bmatrix}. \tag{2.26}$$

where s and s' are initial and final states, respectively with each consisting of $\mathcal{S}$: {Distrust, Trust}. For example, the transition from state Trust to Distrust after a reliable trial has a probability of 0.8809. Estimated initial state probabilities for Distrust and Trust are

$$p_0(\text{Distrust}) = 0.1985 \qquad p_0(\text{Trust}) = 0.8015 \ . \tag{2.27}$$

Adaptive QDA model using Psychophysiological Data

Adaptive implementation of the classification algorithm inherently requires processing the data and estimating trust in real-time. Therefore, we need to continuously extract features from psychophysiological measurements, which is achieved by continuously considering short segments of signals for calculations. We divided the entire duration of the study into multiple 4-second epochs (segments) with 50% overlap between each consecutive epoch. We assume that the decisive cognitive activity occurs after the participant sees the feedback based upon their previous response. Therefore, we only considered the epochs which were in between each successive beginning of a trial and response (trust/distrust) for training the classifier. All epochs were used for prediction. We extracted an *exhaustive set* of potential features from the data for each epoch. We then reduced the dimension of this feature set to include only the statistically significant variables of trust. This reduced feature set was used for classifier modeling and validation.

Feature Extraction: For each of the seven channels (Fz, C3, Cz, C4, P3, POz, and P4) of EEG data, we extracted both frequency and time domain features from each epoch as described in [102]. For frequency domain features, we decomposed each channel's data into four spectral bands, namely delta (0 Hz - 4 Hz), theta (4 Hz - 8 Hz), alpha (8 Hz - 16 Hz), and beta (16 Hz - 32 Hz) and calculated the mean, variance, and signal energy for each band of each epoch. This introduced 84 ($7 \times 4 \times 3$) potential features. For time domain features, we included mean, variance, peak-to-peak values,

mean frequency, root-mean-square, and signal energy of each epoch, thus introducing 42 (7×6) more potential features. Furthermore, to consider the interaction between different regions of the brain, we calculated the correlation between pairs of channels for each epoch, adding another 21 features.

Feature Selection: The EEG data resulted in 147 ($84 + 42 + 21$) potential features. To avoid "the curse of dimensionality" [112], these features were reduced to a smaller feature set using a filter approach feature selection algorithm [141]. Participants were randomly divided into two sets, namely, a training-set consisting of 23 participants and a validation-set consisting of 22 participants. Using only training-set participants' data, we selected the best 15 features using the *Scalar Feature Selection* technique [120, 141]. Fisher Discriminant Ratio (FDR) was used as the class separability criterion with a penalty proportional to the cross-correlation between features. This penalty ensures that the selected features are least correlated, therefore reducing redundancy between features. The selected features are shown in Table 2.11.

Table 2.11.
Features used as input variables for trust classification

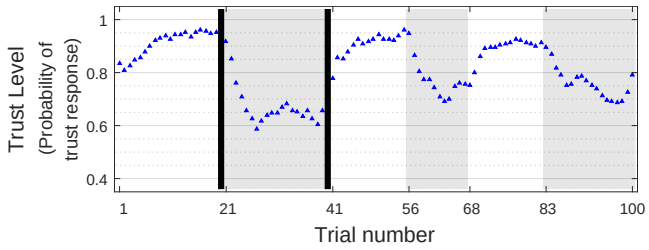
	Feature	Domain
1	Mean Frequency - P4	Time
2	Mean Frequency - C4	Time
3	Mean Frequency - P3	Time
4	Peak-to-peak - C4	Time
5	Peak-to-peak - C3	Time
6	Root Mean Square - Fz	Time
7	Energy - Fz	Time
8	Variation - Fz	Time
9	Correlation - C4 & P4	Time
10	Energy of Beta Band - P3	Frequency
11	Energy of Beta Band - Cz	Frequency
12	Energy of Beta Band - C3	Frequency
13	Variation of Beta Band - P3	Frequency
14	Variation of Beta Band - Cz	Frequency
15	Variation of Beta Band - C3	Frequency

Modeling and validation: The selected feature set was extracted from EEG data to construct the input $\boldsymbol{x}$ to evaluate $P(\boldsymbol{x}|\mathcal{C}_k)$ using (2.20). It should be noted that for each class label $\mathcal{C}_k$, $\mu_k \in R^{n\times 1}$ and $\Sigma_k \in R^{n\times n}$, where n is the cardinality of the feature set. Therefore, for each class label, $n(n+3)/2$ parameters need to be estimated. This is a relatively large number of parameters given our number of data points. For example, for a two class problem with 15 features, the number of parameters to be estimated is 270 using approximately 270 data points in our study. This leads to significant variations in the estimated covariance matrices and often leads to ill-conditioned matrices which cannot be inverted. This is particularly a challenge during the initial estimation period when even fewer data are available. Therefore, to avoid inversion of ill-conditioned matrices and reduce the number of parameters to be estimated, we assume that the features are independent of each other. This results in covariance matrices that are diagonal and easily invertible. Furthermore, this reduces the number of parameters to be estimated to $2n$ for each class label (i.e. 60 parameters in our example above).

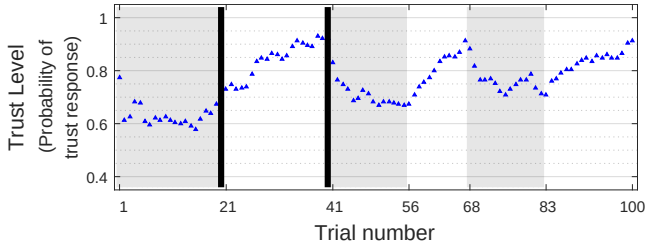
We included psychophysiological measurements in order to identify any latent indicators of trust and distrust. We hypothesized that the trust level would be high in reliable trials and be low in faulty trials, which was validated using responses collected from 581 online participants via Amazon Mechanical Turk [97] as shown in Fig. 2.26 [46]. Therefore, data from reliable trials were labeled as trust, and data from faulty trials were labeled as distrust. In the next section, we use these features extracted from psychophysiological data, along with the dynamic behavioral model derived in Section 2.3.3, to implement the proposed classification algorithm.

2.3.5 Results and Discussions

We implemented the Adaptive Quadratic Discriminant Analysis classifier with Markov Decision Process-based prior probability (hereafter called AQDA-MDP) using the selected features $\boldsymbol{x}$ shown in Table 2.11, class labels $\mathcal{C}_k \in \{\text{Distrust}, \text{Trust}\}$,



(a)



(b)

Figure 2.26. Participants' trust level (blue dots). Subfigure (a) corresponds to group 1 participants and subfigure (b) corresponds to group 2 participants. Faulty trials are highlighted in gray, and black lines mark the breaks between databases.

state transition matrix as given in (2.26), and the initial state probability as given in (2.27). For comparison, we also consider the Adaptive Quadratic Discriminant Analysis classifier (hereafter, called AQDA) exclusively with the prior probability estimated using (2.23). The forgetting factor λ was taken as 1, i.e., no forgetting was used. The algorithms were used for online training and validation of trust classification models from the real-time data for each participant individually.

The results for two different training-set participants and for two different validation-set participants are shown in Fig. 2.27 and Fig. 2.28, respectively. Faulty trials are highlighted in gray, and reliable trials are highlighted in white. A high probability of trust is expected in reliable trials, and a low probability of trust is expected in faulty trials. To observe the benefits of adaptation and to compare the performance of each models, we calculate the mean trial accuracy for each trial. Mean trial accuracy is calculated as the average, across participants, of the percentage of correct prediction

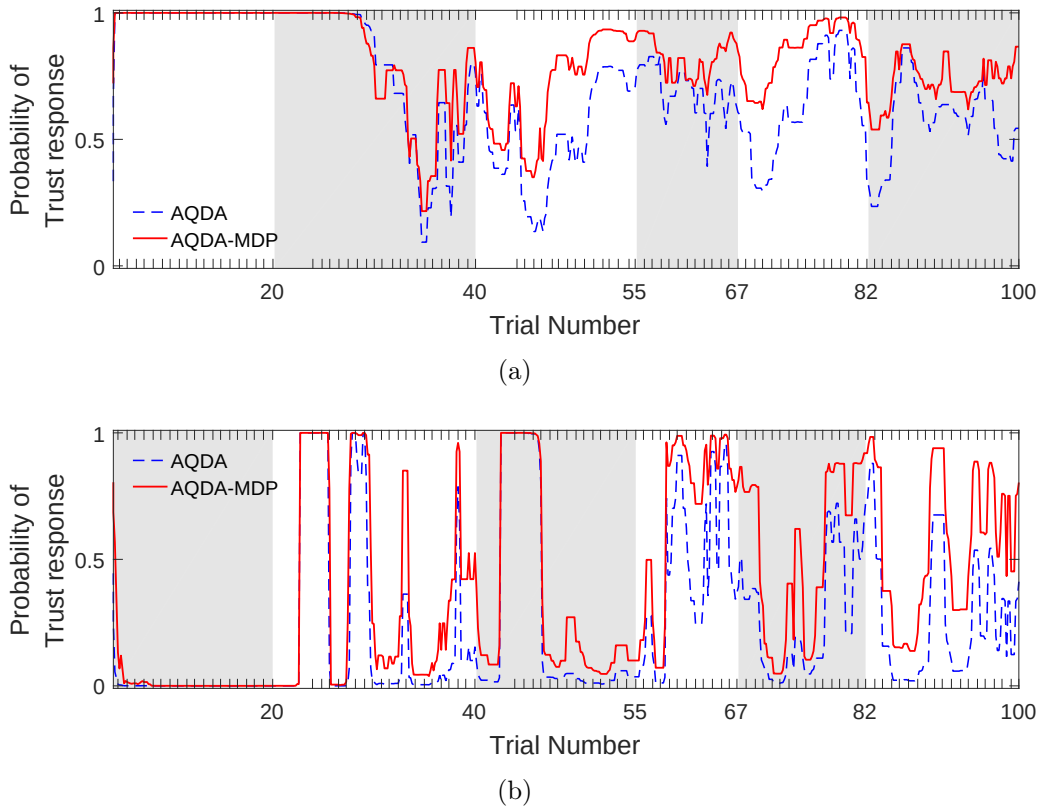
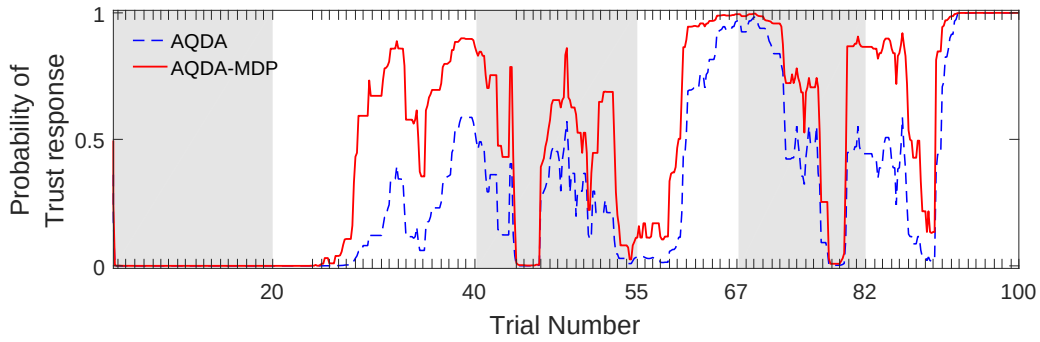


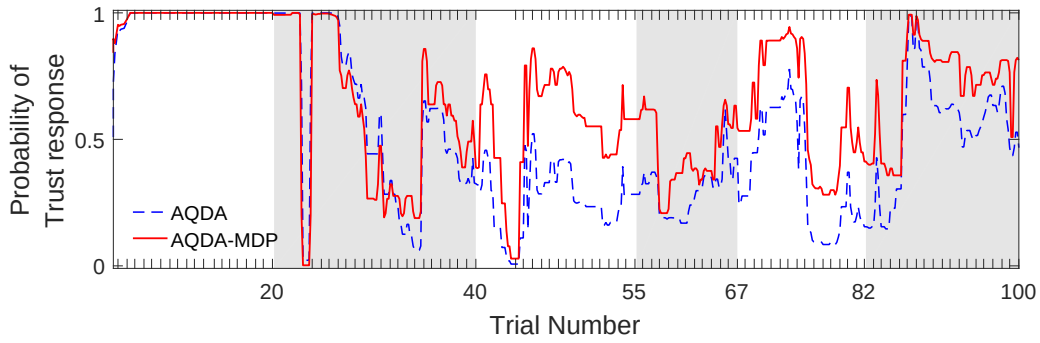
Figure 2.27. Training-set participants' trust level predictions using AQDA-MDP and AQDA algorithms. The top figure (a) shows the prediction of trust for participant 5 in the training set. The bottom figure (b) shows the prediction of trust for participant 7 in the training set. Faulty trials are highlighted in gray.

for epochs for each trial. The variation of mean trial accuracy for training-set and validation-set participants are shown in Fig. 2.29(a) and Fig. 2.29(b), respectively. It can be seen that the performance of the classifier is consistent between training-set and validation-set participants. Therefore, the selected set of features are capable of predicting trust behavior.

We see that the accuracy of the classifier is high for the first 20 trials (see Fig. 2.29). This is the consequence of the experiment design, which has data for one of the classes (either trust or distrust) initially, therefore making the classifier biased towards the initial training data. Consequently, the classifier accuracy just after the 20^{th} trial is



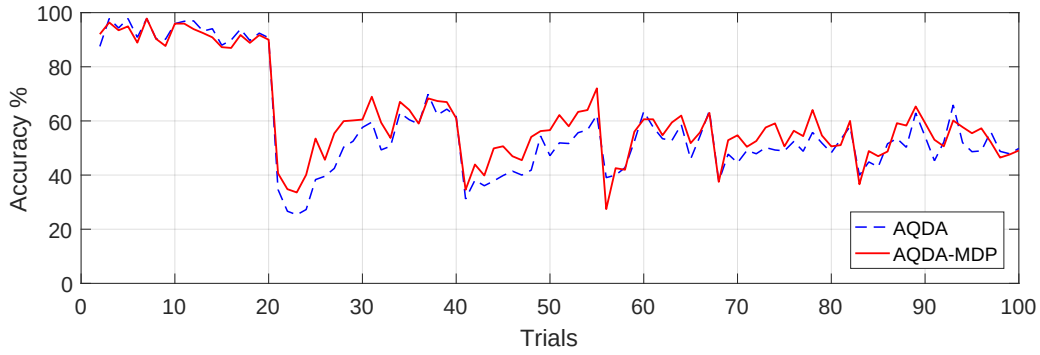
(a)



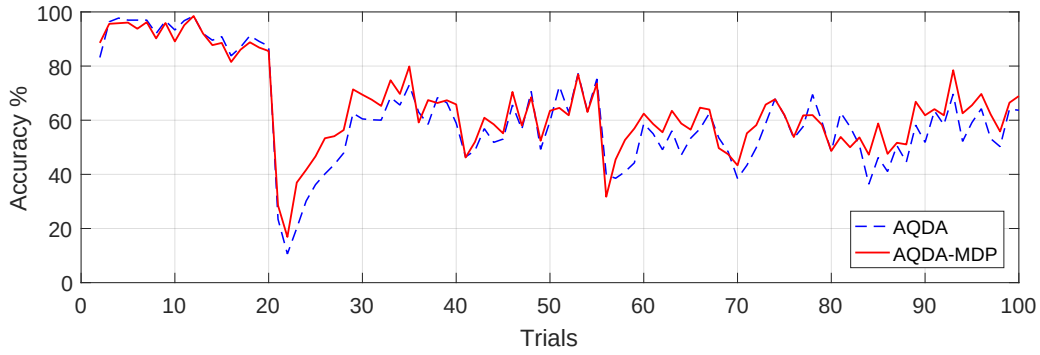
(b)

Figure 2.28. Validation-set participants' trust level predictions using AQDA-MDP and AQDA algorithms. The top figure (a) shows the prediction of trust for participant 36 in the validation set. The bottom figure (b) shows the prediction of trust for participant 34 in the validation set. Faulty trials are highlighted in gray.

poor, and it takes approximately 4-5 trials to eliminate the bias effect and have a considerable sample size for both classes. After the 55^{th} trial, the classifier prediction accuracy decreases as shown in Fig. 2.29. One of the potential reasons is improper class labeling of the data. We assumed that the participants trusted the obstacle detection sensor during the reliable trials and distrusted it during the faulty trials. However, in the later trials during which the sensor reliability changes more rapidly, participants may have been unsure about the system performance. Therefore, our assumption for class labeling may not hold for data collected during these trials. As a result, the adaptive algorithm incorrectly trains itself in the later trials, resulting in



(a)



(b)

Figure 2.29. Mean Trial accuracy for ADQA and AQDA-MDP algorithms. Subfigure (a) corresponds to training-set participants and subfigure (b) corresponds to validation-set participants.

accuracy approximately between 40% and 65% as shown in Fig. 2.29. A better way to label the trials as trust or distrust could improve the performance of the classifier and is the subject of future work. The mean trial accuracy for AQDA-MDP is, in general, higher than that of AQDA. Despite the limitations of class labeling for our experiment, the proposed algorithm enables the combination of two different types of modeling frameworks, a static QDA classifier and a dynamic MDP, systematically using a Bayesian approach to yield a classifier with improved accuracy. More generally, this algorithm can be used for classification of other human behaviors measured using

psychophysiological data and behavioral responses, as well as other dynamic processes characterized by data with non-stationary distributions.

2.4 Chapter Summary

In this chapter, we presented multiple approaches to model and estimate human trust based on human behavioral responses and psychophysiological measurements. We first established a quantitative trust model based on human behavior, motivated by literature on computational models and parameterized using human subject data. This model was verified using data collected from over 800 participants and has a prediction accuracy higher than 92% for the general population. We introduced the effect of cumulative trust, expectation bias, and misses/false alarms, to accurately capture human trust dynamics during human-machine interactions. While the state-space trust model is representative of a population of individuals rather than trained to a specific human, such a model could be used to design machines that are required to interact with unspecified users grouped by demographics.

However, in some uncertain and unstructured environments, it may not be practical to retrieve human behavior continuously for use in a feedback control algorithm. Therefore, for these scenarios, we developed a classification model to estimate human trust using an individual's psychophysiological measurements, specifically EEG and GSR. These models used human subject data collected from 45 participants. The first approach was to consider a common set of psychophysiological features as the input variables for any human and train a classifier-based model using this feature set, resulting in a general trust sensor model with a mean accuracy of 71.22%. The second approach was to consider a customized feature set for each individual and train a classifier-based model using that feature set; this resulted in a mean accuracy of 78.55%. While it is expected that using a feature set customized to a particular individual will outperform a model based upon the general feature set, the time needed for training such a model may be prohibitive in certain applications.

A limitation of the proposed classification algorithm is that it does not consider the temporal dynamics of human behavior and the non-stationary characteristics of psychophysiological signals. Therefore, we described an adaptive probabilistic classification algorithm which uses a dynamic MDP model to incorporate these temporal dynamics. First, we estimated the parameters for a MDP using behavioral responses. We then extracted an exhaustive set of features from psychophysiological data from 23 training-set participants and reduced the dimension of the feature space using scalar feature selection. We trained a real-time adaptive QDA-based classifier using data collected online for these 23 participants. The classifiers were validated against human subject data from another 22 validation-set participants, and an improved estimation accuracy was achieved using the classifier augmented with a dynamic MDP.

For all of the modeling approaches, to elicit the dynamics of trust, we varied automation reliability and then modeled and estimated the resulting trust dynamics. However, although automation reliability strongly affects human trust, and the modeled relationship can be used to *predict* trust, automation reliability should not be considered a control variable for the purpose of *affecting* human trust. As discussed in Chapter 1, automation transparency can instead be varied to affect human trust and human workload. Therefore, we model the dynamic effects of an automation's transparency on human trust and workload in the next chapter so that it can be used as a control variable for improving human-machine collaboration.

3. TRANSPARENCY-BASED FEEDBACK CONTROL OF HUMAN TRUST

The contents of this chapter have been accepted for publication in the *IEEE Control Systems Magazine* [142] and are reported here with minor modifications.

3.1 Introduction

Published studies have shown that human trust in automation is an important factor that affects the outcome of the interactions and that it can be improved by increasing the transparency of an automation's decisions [8, 9]. Chen et al. (2014) defines transparency as "the descriptive quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process." [7] Therefore, greater transparency allows humans to make informed judgments and accordingly make better choices. In this chapter, we model the dynamic effects of an automation's transparency on human behavior and use it as a control variable for improving human-machine collaboration.

High levels of trust are not always desirable and can lead to humans trusting an error-prone system. Instead, trust should be appropriately calibrated according to the system's capability [4]. Moreover, high transparency involves communicating more information to the human and thus can increase the workload of the human [10]. In turn, high levels of workload can lead to fatigue, which can reduce the human's performance. Therefore, we aim to design intelligent systems that can respond to changes in human trust and workload in real-time to achieve optimal or near-optimal performance. For intelligent systems, a user interface (UI) is generally the means through which communication with the human is achieved. Therefore, the system

must understand how the *transparency* of its communication through the UI affects the human's cognitive state.

Although researchers have developed various models of human trust behavior [44, 45] and established the effect of transparency on trust [8, 9, 23], there does not exist a quantitative model that captures the *dynamic* effect of transparency on human trust. Furthermore, published studies considering the effects of transparency on workload do not model its dynamics. Therefore, a fundamental gap remains in capturing the dynamic effect of machine transparency on human trust-workload behavior so that it can be used for improving human-machine collaboration.

In this chapter, we present a partially observable Markov decision process (POMDP) model framework for capturing *dynamics of human trust and workload* for contexts that involve interaction between a human and an intelligent decision-aid system. We specifically consider a reconnaissance mission study adapted from the literature in which human subjects are aided by a virtual robotic assistant in completing a series of reconnaissance missions. We use the collected human subject data to train the POMDP model. We further study the effects of transparency and experience on human trust and workload using the estimated parameters. In Section 3.4, the trained model is used to estimate human trust and workload and to develop a near-optimal control policy that varies machine transparency to improve outcomes of the human-machine collaboration.

3.2 Modeling Human Trust and Workload

Researchers have developed various models for human trust. Qualitative models [143–146] are useful for defining which variables affect trust but are insufficient for making quantitative predictions. On the other hand, regression models [147, 148] quantitatively capture trust but do not consider its dynamic response characteristics. To fill this gap, researchers have proposed both deterministic models [4, 34, 37, 46, 47, 59, 149, 150] and probabilistic models [143, 151, 152] of human trust dynamics. With

respect to probabilistic approaches, several researchers have modeled human trust behavior using Markov models, particularly hidden Markov models (HMMs) [44,45,153]. While HMMs can be used for intent inference and to incorporate human behavior related uncertainty [154–157], they do not include the effects of inputs or actions from autonomous systems that affect human behavior. On the other hand, models based on Markov decision processes (MDP) do consider the effect of inputs or actions and have been used to model human behavior for human-in-the-loop control synthesis [158]. However, MDPs do not account for the unobserved nature of human cognitive constructs like trust and associated uncertainties. A useful extension of HMMs and MDPs, called partially observable Markov decision processes (POMDPs), provides a framework that accounts for actions/inputs as well as unobserved states and also facilitates calculating the optimal series of actions based on a desired reward function. Recent work has demonstrated the use of a POMDP model with human trust dynamics to improve human-robot performance [159]. POMDPs have also been used in HMI for automatically generating robot explanations to improve performance [23–25] and estimating trust in agent-agent interactions [160]. For example, the POMDP model in [23–25] is used to simulate only the dynamics of the robot's decisions and generates recommendations of different transparency levels. However, the model does not capture human trust-workload behavior nor the dynamic effects of automation transparency on that behavior. In this work, we model the human trust-workload behavior as a POMDP and optimally vary automation transparency to improve human-machine interactions.

Trust and workload levels of humans have been classically obtained using self-reported surveys. Trust surveys involve questions customized to an experiment along with a Likert scale for the participants to report how much they trusted the system and understood the scenario [145]. Workload is commonly assessed using the NASA TLX survey [161]. In the context of real-time feedback algorithms, however, it is not practical to use surveys for human measurements because continuously inquiring humans is generally not feasible. Alternatively, we can use behavioral metrics that

are readily available in real time and correlate to human trust-workload behavior, including compliance and response time. Compliance is defined as the human agreeing to the automation's recommendation when one is issued. Human response time is the time a human takes to respond to a stimulus [162]. These metrics can be implicitly used to infer the underlying trust and workload states of the human.

Several studies have shown a strong correlation between human trust and compliance. For example, researchers have shown that perception of trust is associated with improved compliance [163]. Furthermore, studies showed that trust and compliance exhibited similar patterns with variations in system accuracy [164, 165]. Studies in [24] also confirmed the correlation between trust and compliance during human-robot interaction. Similarly, other studies have shown a correlation between workload and response time [9, 29]. The peripheral detection task (PDT) method based on human response time has been shown to be a sensitive measure of cognitive workload [166, 167]. Newell and Mansfield showed that with environmental stressors, as participants' reaction times slowed down, simultaneously their workload demands based on the NASA TLX assessment also increased [168]. Therefore, in this work, we assume a causal relationship of trust affecting compliance and propose to use response times as observations corresponding to workload.

3.2.1 POMDP Model of Human Trust and Workload

Here we consider contexts that involve human interaction with a decision-aid system that gives recommendations based on the presence or absence of a stimulus. During such an interaction, the final decision and action is taken by the human; the decision-aid system only provides a *recommendation* to the human. Although such systems are a subset of autonomous systems, they are widely used in safety-critical situations, such as assistive robots used for threat detection in the military theater or health recommender systems used for detecting diseases. When interacting with a decision-aid system, the human is typically choosing to either comply with, or

reject, the system's recommendation. This human decision has an associated response time (RT). Since we cannot directly observe human trust and workload states, we use human observations—compliance and response time—to estimate the states. We further assume that human trust and workload are influenced by characteristics of the decision-aid system's recommendations. In particular, we consider the effects of the system's recommendation and transparency, and the human's past experience with the system. Also, the previous states of trust and workload affect the current state. Therefore, with an assumption that the dynamics of human trust and workload follow the Markov property [169], we use a POMDP to model the human trust-workload behavior.

A Partially Observable Markov Decision Process (POMDP) is an extension of a Markov decision process (MDP) that accounts for partial observability through hidden states. It is similar in structure to the classic discrete-time state-space model as described in Table 3.1 but with a discrete state-space. Formally, a POMDP is a 7-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{E}, \mathcal{R}, \gamma)$ where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, and $\mathcal{O}$ is a set of observations and can be represented as shown in Figure 3.1. The transition probability function $\mathcal{T}(s'|s, a)$ governs the transition from the current state s to the next state s' given the action a. The emission probability function $\mathcal{E}(o|s)$ governs the likelihood of observing o given the process is in state s. Finally, the reward function $\mathcal{R}(s', s, a)$ and the discount factor γ are used for finding an optimal control policy. A detailed description of MDPs and POMDPs can be found in [170].

We define the finite set of states $\mathcal{S}$ consisting of tuples containing the *Trust* state s_T and the *Workload* state s_W, respectively, where each state can take on a low ($\bullet_\downarrow$) or high ($\bullet_\uparrow$) value. The characteristics of the system recommendations are defined as the finite set of actions $\mathcal{A}$, consisting of tuples containing *Recommendation*, *Experience*, and *Transparency*. Here, *Recommendation* of the automation a_{S_A} can be either Stimulus Absent S_A^- or Stimulus Present S_A^+; *Experience* a_E, which depends on the reliability of the last recommendation, can be either Faulty E^- or Reliable E^+; and *transparency* a_τ can be either Low Transparency τ_L, Medium Transparency τ_M, or

Table 3.1.
Similarities between a Partially Observable Markov Decision Process (POMDP) and a discrete-time state-space model.

	POMDP	State-space model
States	$s \in \mathcal{S}$	$x \in \mathbb{R}_x$
Actions/Inputs	$a \in \mathcal{A}$	$u \in \mathbb{R}_u$
Observations/Outputs	$o \in \mathcal{O}$	$y \in \mathbb{R}_y$
Transition function	$p(s') = \mathcal{T}(s'\|s, a)$	$x_{t+1} = f(x_t, u_t)$
Emission/Output function	$p(o) = \mathcal{E}(o\|s)$	$y_t = g(x_t)$
Reward/Cost function	$\mathcal{R}(s', s, a)$	$\mathcal{L}(x_t, u_t)$
Optimal control policy (a^*/u^*)	$\underset{a\in\mathcal{A}}{\text{argmax}} \sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s^{t+1}, s^t, a^t)$	$\underset{u\in\mathbb{R}_u}{\text{argmin}} \sum_{t=0}^{\infty} \mathcal{L}(x_t, u_t)$

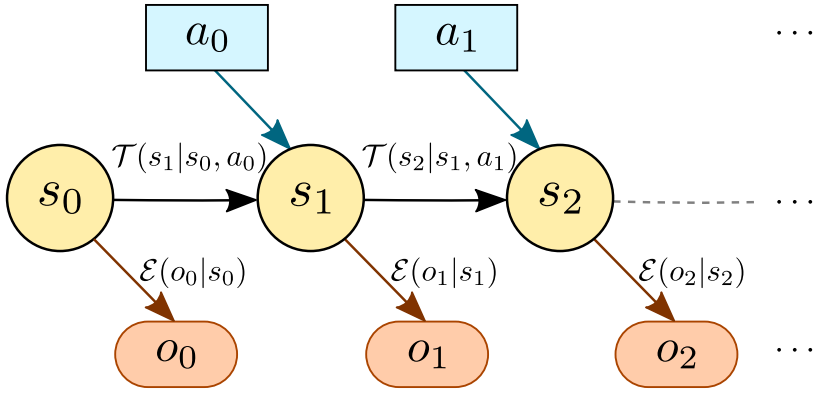


Figure 3.1. A simplified representation of a partially observable Markov decision process (POMDP) model.

High Transparency τ_H. The three levels of transparency depend on the context and the automation. The observable characteristics of the human's decision are defined as the set of observations $\mathcal{O}$ consisting of tuples containing *compliance* and *Response Time*. Here, *Compliance* o_C can be either Disagree C^- or Agree C^+ and *Response Time* $o_{RT} \in \mathbb{R}^+$ is defined as the time the human takes to respond after receiving the decision-aid's recommendation. The definition of our trust-workload POMDP model is summarized in Table 3.2.

Table 3.2.
Definition of the trust-workload POMDP model. Human trust and workload are modeled as hidden states that are affected by actions corresponding to the characteristics of the decision-aid's recommendations. The observable characteristics of the human's decisions are modeled as the observations of the POMDP.

States $s \in \mathcal{S}$	$s = \begin{bmatrix} \textit{Trust } s_T \\ \textit{Workload } s_W \end{bmatrix}$	$s_T \in T$ $T = \begin{Bmatrix} \text{Low Trust } T_\downarrow, \\ \text{High Trust } T_\uparrow \end{Bmatrix}$
		$s_W \in W$ $W = \begin{Bmatrix} \text{Low Workload } W_\downarrow, \\ \text{High Workload } W_\uparrow \end{Bmatrix}$
Actions $a \in \mathcal{A}$	$a = \begin{bmatrix} \textit{Recommendation } a_{S_A} \\ \textit{Experience } a_E \\ \textit{Transparency } a_\tau \end{bmatrix}$	$a_{S_A} \in S_A$ $S_A = \begin{Bmatrix} \text{Stimulus Absent } S_A^-, \\ \text{Stimulus Present } S_A^+ \end{Bmatrix}$
		$a_E \in E$ $E = \begin{Bmatrix} \text{Faulty last experience } E^-, \\ \text{Reliable last experience } E^+ \end{Bmatrix}$
		$a_\tau \in \tau$ $\tau = \begin{Bmatrix} \text{Low Transparency } \tau_L, \\ \text{Medium Transparency } \tau_M, \\ \text{High Transparency } \tau_H \end{Bmatrix}$
Observations $o \in \mathcal{O}$	$o = \begin{bmatrix} \textit{Compliance } o_C \\ \textit{Response Time } o_{RT} \end{bmatrix}$	$o_C \in C$ $C = \begin{Bmatrix} \text{Disagree } C^-, \\ \text{Agree } C^+ \end{Bmatrix}$
		$o_{RT} \in \mathbb{R}^+$

We assume that human trust and workload behavior are conditionally independent given an action. Furthermore, we assume that trust only affects compliance, and workload only affects response time. This enables the trust and workload models to be identified independently. Moreover, it significantly reduces the number of parameters in each model and in turn, the amount of human data needed for training each model.

The transition probability functions for the trust model, $\mathcal{T}_T : T \times T \times \mathcal{A} \rightarrow [0, 1]$, and for the workload model, $\mathcal{T}_W : W \times W \times \mathcal{A} \rightarrow [0, 1]$, are represented by $2 \times 2 \times 12$ matrices mapping the probability of transitioning between the states of trust $s_T \in T$ or workload $s_W \in W$ after an action $a \in \mathcal{A}$. For the trust model, the emission probability function $\mathcal{E}_T : C \times T \rightarrow [0, 1]$ is represented by a 2×2 matrix, mapping the probability of observing Compliance $o_C \in C$ given the state of trust s_T. Similarly, for the workload model, the emission probability function $\mathcal{E}_W : \mathbb{R}^+ \times W \rightarrow [0, 1]$ is represented by two probability density functions, each representing the probability of observing a response time $o_{RT} \in \mathbb{R}^+$ given the state of workload s_W. Human reaction-time has been shown to have a distribution similar to the ex-Gaussian distribution [162, 171, 172], which is a convolution (mixture) of a Gaussian and an exponential distribution. Here we assume that each workload state has a characteristic response time distribution defined by an ex-Gaussian distribution.

Human response time RT (also called reaction time or latency [172]) is the time duration between the presentation of stimulus to a human and the human's response [162]. Response time analysis has a long history in experimental psychology and still is used as a dominant dependent measure to identify the processes that affect the human's response. Statistically, RT is often treated as a random variable because it typically varies between trials for the same human within a given context. Furthermore, RT distributions are attributed with a positively skewed unimodal shape (see Figure 3.2) that cannot be effectively captured by only mean and variance [173]. Therefore, a RT distribution cannot be modeled as a Gaussian distribution.

To describe a typical RT distribution, standard distributions like gamma distribution and log-normal distribution have been used in the literature [174, 175]. One of the most widely employed distribution for RT data has been an exponentially modified Gaussian (or ex-Gaussian) distribution [171, 176–182], defined as the convolution of an exponential distribution with a Gaussian distribution. This distribution is characterized by three parameters, μ, σ and τ, with μ and σ characterizing the average and standard deviation of the Gaussian component and τ characterizing the decay

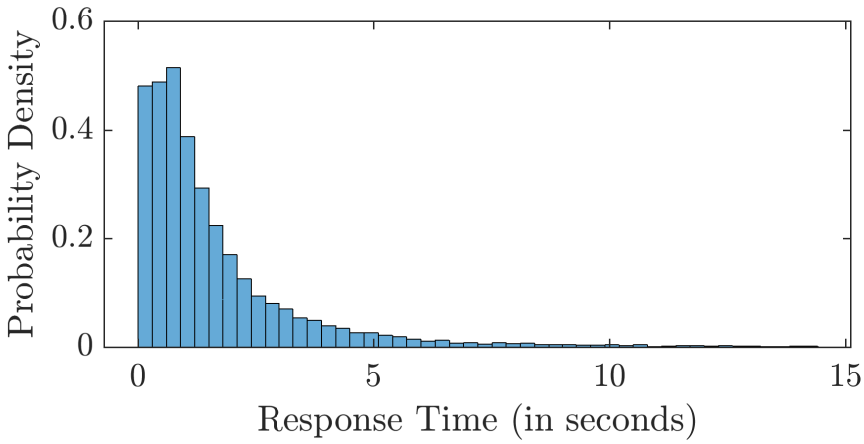


Figure 3.2. Empirical probability density function representing the response time RT distribution for the aggregated human subject study data described in Section 3.2.2. RT distributions are attributed with a positively skewed unimodal shape with a rapid rise on the left and a long positive tail on the right.

rate of the exponential component. For σ and τ greater than zero, the probability density function for the ex-Gaussian distribution is

$$f(x) = \frac{1}{2\tau} \exp\left(\frac{\sigma^2}{2\tau^2} - \frac{x-\mu}{\tau}\right) \operatorname{erfc}\left(\frac{\sigma^2}{\tau} - (x-\mu)\right) \quad ,$$

where erfc is the complementary error function defined as

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_0^\infty e^{-t^2} dt \quad .$$

Researchers have attempted to specify the underlying process for RT that leads to the ex-Gaussian distribution by attributing the exponential component to decision processes and the normal component to residual processes [183]. However, this rationale still remains unproven [184]. Nonetheless, the ex-Gaussian distribution has been found to fit the RT distribution better than gamma and log-normal distributions.

3.2.2 Human Subject Study

To parameterize the human trust and workload models, we collect human subject data in a specific decision-aid system context. The experiment presented here is adapted from an earlier study [25]. The modified experiment captures the effects of different levels of system transparency on human trust and workload behavior along with the human-robot interaction performance.

Stimuli and Procedure: A within-subjects study was performed in which participants were asked to interact with a simulation consisting of multiple reconnaissance missions. Each participant performed three missions while being assisted by a decision-aid robot. In each mission, participants were required to search 15 buildings and mark them as safe or unsafe based on the presence or absence of gunmen. The goal of each mission was to search all of the buildings as fast as possible. Prior to entering each building, the participant needed to decide if they would use light armor or heavy armor while searching the building. They were informed that searching a building with heavy armor would take approximately 7 seconds but would ensure that they would not be injured if gunmen were present. On the other hand, searching with light armor would take only 3 seconds, but if gunmen were present, the participant would be injured and penalized with a 20-second recovery time. To assist the participant, the decision-aid robot would survey each building first and make a recommendation on which armor to use.

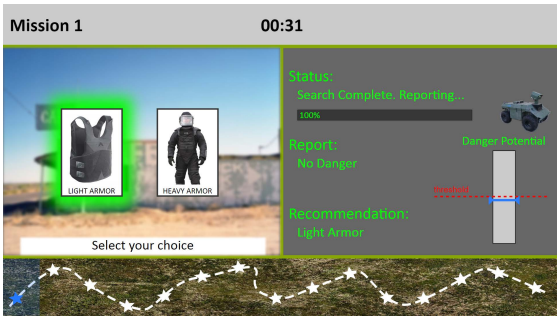
In each mission, the participant was assisted by a robot that used a different transparency level for its recommendation. The interface for each of the levels of transparency is shown in Figure 3.3. The low transparency robot reported if gunmen were present or absent along with the armor recommendation (see Figure 3.3(a)). The medium transparency robot additionally included a sensor bar indicating the level of potential danger as perceived by the robot (see Figure 3.3(b)). The sensor reading was below the robot's threshold when no gunmen were detected and above the threshold when gunmen were detected. The high transparency robot included

all of the information provided by the medium transparency robot, along with seven thermal images collected from inside the building (see Figure 3.3(c)). Note that this is only one way of defining different levels of transparency and can vary based on feasibility, context, and automation.

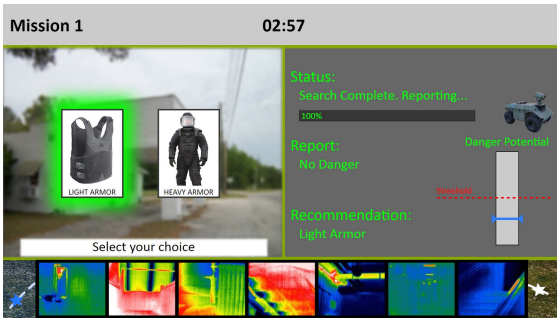
Before the participants began the actual mission, they completed a tutorial mission consisting of six trials that helped familiarize them with the study interface and the three levels of transparency. The tutorial mission was uniform across all participants. For the experiment itself, the order of missions for each transparency level was randomized across participants to reduce ordering effects [185]. This randomization reduces the impact of factors like experience, practice from previous missions, and fatigue on the analysis. The presence or absence of gunmen was equally probable in each trial. The robots' recommendations were 70% accurate. When the robot's recommendation was incorrect, it was a false alarm (false positive) or miss (false negative) with equal probability. The sequence of events in each trial is shown in Figure 3.4.

Participants: Two hundred and twenty-one participants from the United States participated in, and completed, the study online. They were recruited using Amazon Mechanical Turk [97], with the criteria that they must live in the US and have completed more than 1000 tasks with at least a 95% approval rate. The compensation was $1.50 for their participation, and each participant electronically provided their consent. The Institutional Review Board at Purdue University approved the study. Since the participants were not monitored while completing the study, we suspect that some participants were not sufficiently engaged with the study, reflected by their unusually high response times to stimuli. Therefore, we filtered data from participants who had any response time longer than the threshold at 99.5 percentile of all response times, which was approximately 40.45 seconds. As a result, 25 outlying participants were removed from the dataset.

(a)



(b)



(c)

Figure 3.3. Example screenshots of robot reports corresponding to the three levels of transparencies. The top screenshot (a) shows a low transparency case with the robot's report (Gunmen Present) along with the armor recommendation (Heavy Armor). The middle screenshot (b) shows a medium transparency case that additionally includes a sensor bar on the left that indicates the level of potential danger perceived by the robot. The bottom screenshot (c) shows a high transparency case that further includes seven thermal images collected from inside the building, which the human can evaluate themselves.

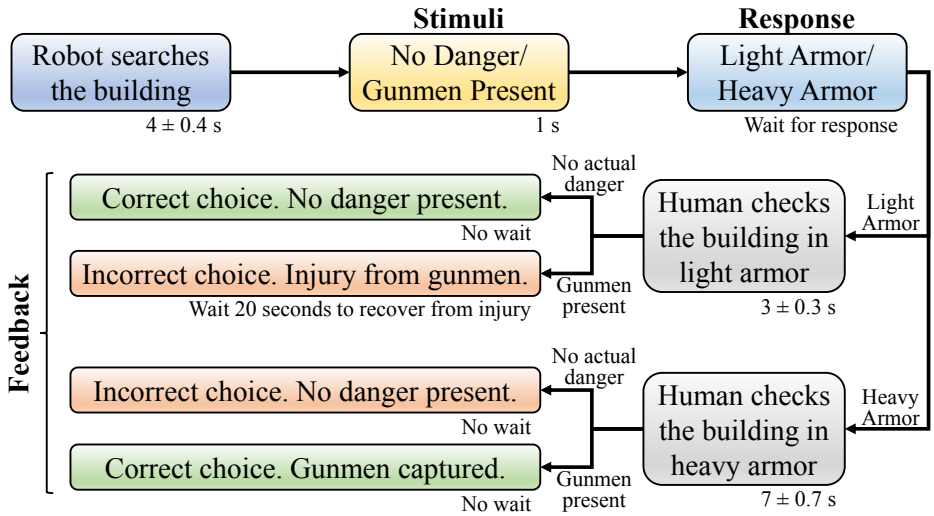


Figure 3.4. The sequence of events in a single trial. The time length marked on the bottom right corner of each event indicates the time interval for which the information appeared on the computer screen.

3.3 Model Parameter Estimation

We assume that the trust and workload behavior of the general population can be represented by a common model. Therefore, we used the aggregated data from all participants to estimate the transition probability function, observation probability function, and the prior probabilities of states for the trust and workload models. For this study, the system recommendation that indicates *Light Armor* is defined as Stimulus Absent S_A^- and the recommendation that indicates *Heavy Armor* is defined as Stimulus Present S_A^+. We define a sequence of action-observation data for a participant as the interaction between the participant and robot in each mission. Therefore, we have 196×3 sequences of data to estimate the parameters of each model.

The problem of model parameter estimation for POMDP models using sequences of data is defined as finding optimal parameters that maximize the likelihood of observing the sequences of observation for the given sequences of actions. For estimating the parameters of the discrete observation-space trust model, we use an extended version of the Baum-Welch algorithm, which is typically used for hidden Markov model

(HMM) estimation (see [186] for details). However, the continuous non-Gaussian distribution of the emission probability function of the workload model makes it infeasible to be estimated using the Baum-Welch algorithm. Therefore, we implement a genetic algorithm using Matlab to optimize the parameters for the workload model in which the algorithm aims to find a set of parameters that maximize the likelihood of the sequences given the model parameters. The likelihood of the sequences is calculated using the *forward algorithm* [186]. Given the model parameters, the forward algorithm computes the joint probability of a state s_k at time k, observations until time k (i.e., $o_{1:k}$), and actions until time k (i.e., $a_{1:k}$), that is, $p(s_k, o_{1:k}, a_{1:k})$, recursively over time by taking advantage of the conditional independence. The likelihood of the sequence is then calculated as the sum of $p(s_N, o_{1:N}, a_{1:N})$ across all states at the end of the sequence at time N, therefore giving the probability of the action-observation sequence $p(o_{1:N}, a_{1:N})$. The forward algorithm reduces the computational complexity of this evaluation from $\mathcal{O}(Nn_s^N)$, if we use the ad hoc method of marginalizing over all possible state sequences, to $\mathcal{O}(Nn_s^2)$, where n_s is the number of states and N is the length of the sequence. The estimated POMDP models of trust and workload models are presented and analyzed in the next section.

3.3.1 Trust Model

The estimated trust model consists of initial state probabilities $\pi(s_T)$, an emission probability function $\mathcal{E}_T(o_C|s_T)$, and a transition probability function $\mathcal{T}_T(s'_T|s_T, a)$. Based on the emission probability function for trust $\mathcal{E}_T(o_C|s_T)$, we define the High Trust state $s_T = T_\uparrow$ as that in which there is a higher probability of observing the human comply with the automation's recommendation, $o_C = C^+$. The estimated initial probabilities of Low Trust $T_\downarrow$ and High Trust $T_\uparrow$ are $\pi(T_\downarrow) = 0.1286$ and $\pi(T_\uparrow) = 0.8714$, respectively. This is consistent with findings that recent widespread use of automation has led to humans trusting a system when they have no experience with it [187]. The emission probability function $\mathcal{E}_T(o_C|s_T)$ is depicted in Figure 3.5

and characterizes the probability of a participant's compliance with the system's recommendations given the participant's state of trust. Both states, Low Trust and

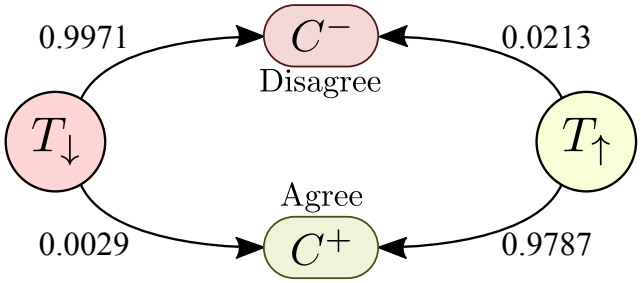


Figure 3.5. Emission probability function $\mathcal{E}_T(o_C|s_T)$ for the trust model. Probabilities of observation are shown beside the arrows. Low Trust has a 99.71% probability of resulting in participants disagreeing with the recommendation and High Trust has a 97.87% probability of resulting in participants agreeing with the recommendation.

High Trust, have more than 97% probability to result in participants disagreeing and agreeing with the recommendation, respectively. However, there is still a small probability of participants disagreeing while in a state of High Trust as well as participants agreeing while in a state of Low Trust. This inherently captures the uncertainty in human behavior.

Figure 3.6 represents the transition probability function $\mathcal{T}_T(s'_T|s_T, a)$ showing the probability of transitioning from the state s_T to s'_T (where $s_T, s'_T \in T$) given the action $a \in \mathcal{A}$. The cases when the recommendation suggests Light Armor S_A^- can be considered relatively high-risk situations in our context because incorrectly complying with a faulty recommendation—that is, wearing Light Armor in the presence of gunmen—can result in getting injured and a penalty of 20 seconds (see (Figures 3.6(a) and 3.6(b)). On the other hand, the cases when the recommendation suggests Heavy Armor S_A^+ are low-risk situations (Figures 3.6(c) and 3.6(d)) as incorrect compliance only leads to an extra 4 seconds of search time. We observe that the probability of transitioning to High Trust $T_\uparrow$ as well as staying in High Trust $T_\uparrow$ is higher for low-risk situations (Figure 3.6(c) and 3.6(d)) as compared to the corresponding high-risk

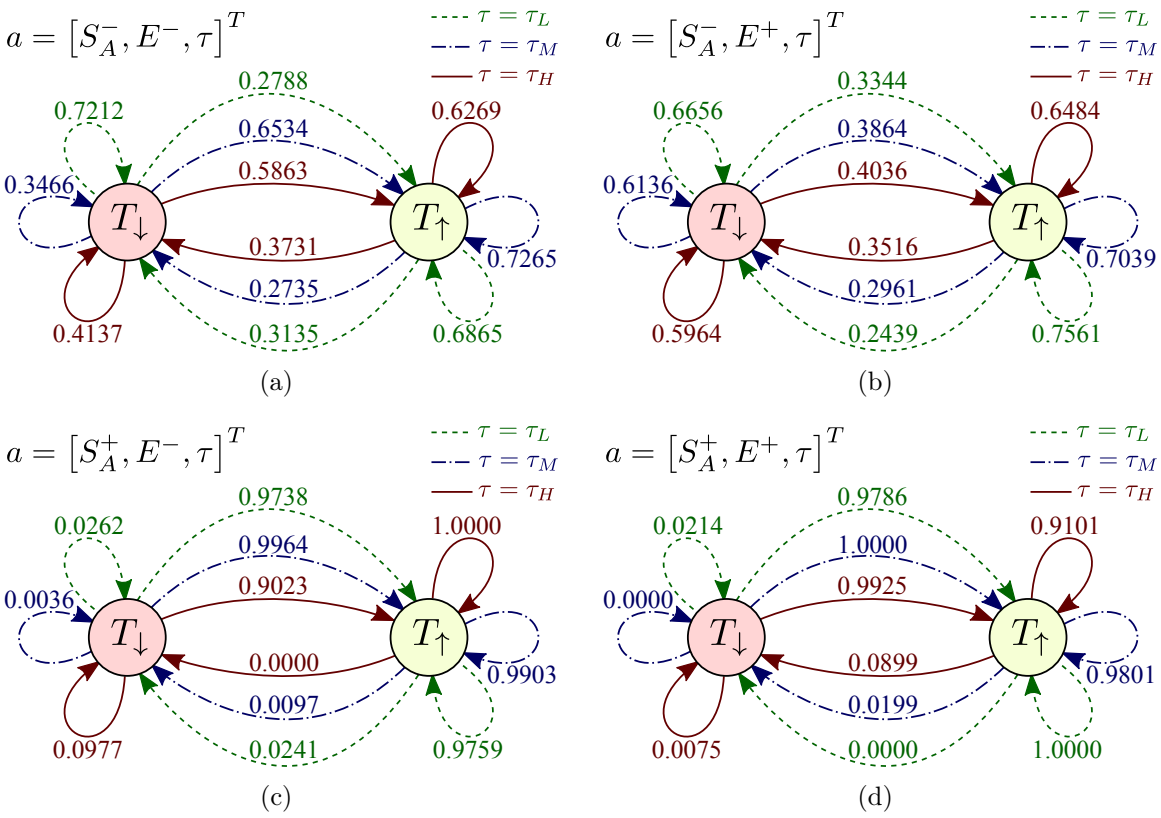


Figure 3.6. Transition probability function $\mathcal{T}_T(s'_T|s_T, a)$ for the trust model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid's recommendation is Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. Both cases (a) and (b) can be considered relatively high-risk situations in this context because incorrectly complying with a faulty recommendation—that is, wearing Light Armor in the presence of gunmen—can result in injury. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

situations (Figure 3.6(a) and 3.6(b)). Given that the robot was only 70% reliable in the study, this over-trust during low-risk situations, with transition probabilities to

High Trust $T_{\uparrow}$ being greater than 91% (see Figure 3.6(c) and 3.6(d)), indicates the inherent conservative behavior of participants. The participants preferred to comply with the robot (by choosing Heavy Armor) and risk an effective penalty of 4 seconds instead of risking a penalty of 20 seconds. It should be noted that the participants did not know about the failure rate of the robot.

Interestingly, we observe that high transparency τ_H has the highest probability of causing a transition from High Trust $T_{\uparrow}$ to Low Trust $T_{\downarrow}$ as compared to lower transparencies in most cases (except Figure 3.6(c)). This is because high transparency enables the participant to make a more informed decision and avoid errors that would result from trusting a faulty recommendation. Therefore, high transparency helps the participant to calibrate their trust correctly in these cases. Moreover, in most cases (except Figure 3.6(c)), low transparency τ_L has the lowest probability of causing a transition from Low Trust $T_{\downarrow}$ to High Trust $T_{\uparrow}$. However, low transparency can offer the best strategy for maintaining a state of high trust (see Figure 3.6(b)). In summary, our findings suggest that transparency does not directly affect human trust; instead, factors such as the human's current trust state, assessment of risk, and the reliability of the automation affect how transparency changes trust.

3.3.2 Workload Model

Similar to the trust model, based on the emission probability function for workload $\mathcal{E}_W(o_{RT}|s_W)$, we define the High Workload state $s_W = W_{\uparrow}$ as that in which the expected response time $\mathbf{E}[o_{RT}|s_W]$ is longer. We estimated the initial probabilities of Low Workload $W_{\downarrow}$ and High Workload $W_{\uparrow}$ to be $\pi(W_{\downarrow}) = 0.3097$ and $\pi(W_{\uparrow}) = 0.6903$, respectively. The high initial probability of High Workload $W_{\uparrow}$ is expected because participants initially need to familiarize themselves with the system. The emission probability function $\mathcal{E}_W(o_{RT}|s_W)$ is represented in Figure 3.7, which shows the probability density functions (PDFs) of observing participants' response time as o_{RT} given their state of workload s_W. We observe that Low Workload $W_{\downarrow}$ is more

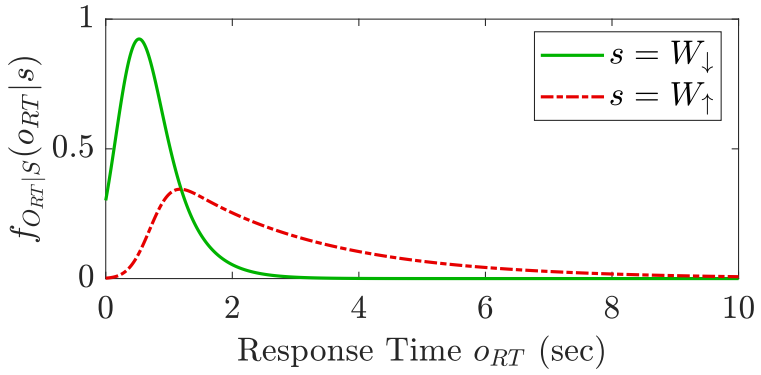


Figure 3.7. Emission probability function $\mathcal{E}_W(o_{RT}|s_W)$ for the workload model. For Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_{\downarrow}}(o_{RT}|W_{\downarrow})$ is characterized by an ex-Gaussian distribution with $\mu_{W_{\downarrow}} = 0.2701$, $\sigma_{W_{\downarrow}} = 0.2964$, and $\tau_{W_{\downarrow}} = 0.4325$. For High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_{\uparrow}}(o_{RT}|W_{\uparrow})$ is characterized by an ex-Gaussian distribution with $\mu_{W_{\uparrow}} = 0.7184$, $\sigma_{W_{\uparrow}} = 0.2689$, and $\tau_{W_{\uparrow}} = 2.2502$. Low Workload $W_{\downarrow}$ is more likely than High Workload to result in a response time of less than approximately 1.19 seconds.

likely than High Workload to result in a response time of less than approximately 1.19 seconds. High Workload is more likely to lead to high response times.

The transition probability function $\mathcal{T}_W(s'_W|s_W, a)$ is represented in Figure 3.8 and shows the probability of a participant transitioning from the state s_W to s'_W based on the action $a \in \mathcal{A}$, where $s_W, s'_W \in W$. We observe that in most cases, the probability of transitioning from Low Workload $W_{\downarrow}$ to High Workload $W_{\uparrow}$ is greater for higher transparencies for a given recommendation and experience (except Figure 3.8(c)). Therefore, it is more likely that higher transparencies will increase participants' workload if they are in a state of Low Workload $W_{\downarrow}$ because they need to process even more information for decision-making. However, if a participant is in a state of High Workload $W_{\uparrow}$, medium transparency τ_M has a lower probability than low transparency τ_L to keep them in a High Workload $W_{\uparrow}$ state. This may occur because in such cases, low transparency may not provide enough information for the participants to make a decision, leaving them confused. This results in participants using more time and

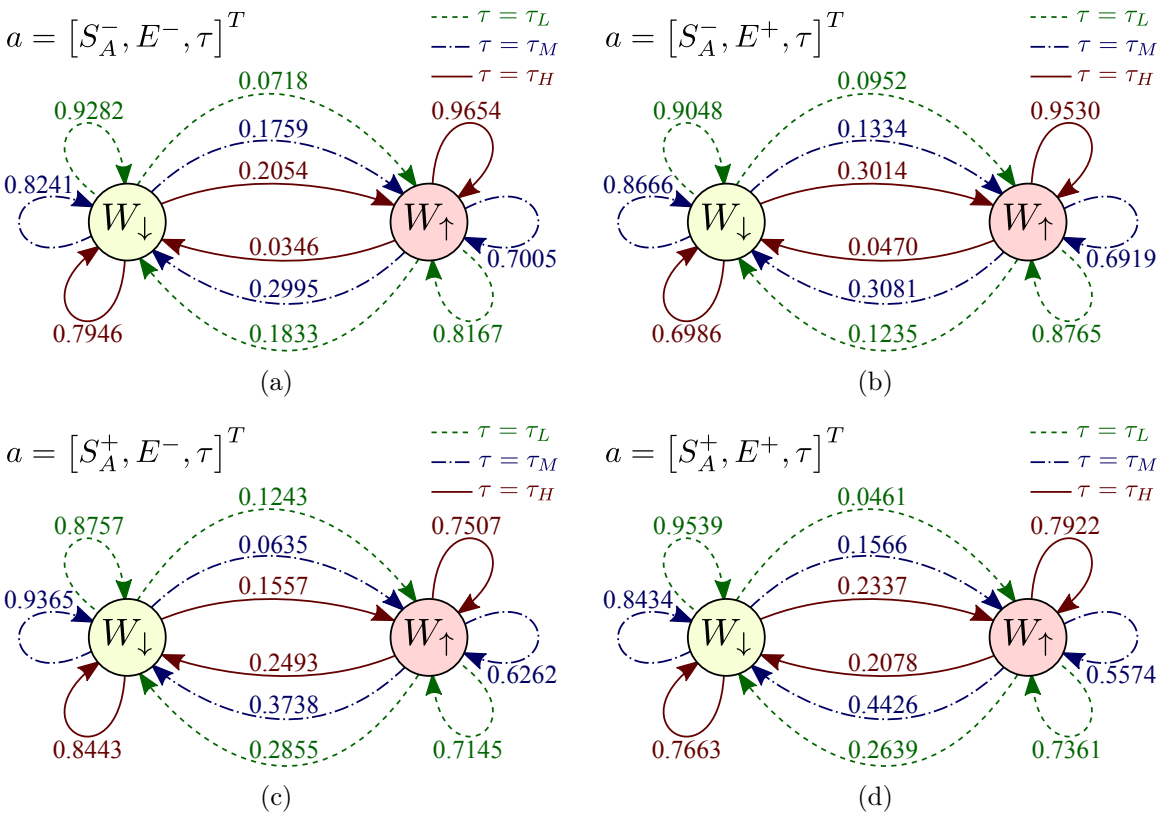


Figure 3.8. Transition probability function $\mathcal{T}_W(s'_W|s_W, a)$ for the workload model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

effort to reach a decision. Overall, high transparency τ_H has the highest probability of keeping the participant in a state of High Workload $W_\uparrow$ for a given recommendation and experience. Finally, it is worth noting that the probability of transitioning to High Workload $W_\uparrow$ from any workload state is higher when the decision-aid rec-

ommends Light Armor S_A^- (see Figure 3.8(a) and 3.8(b)) as compared to when the decision-aid recommends Heavy Armor S_A^+ (see Figure 3.8(c) and 3.8(d)) for a given experience and transparency. This is because a recommendation suggesting Light Armor S_A^- has a higher risk as discussed previously, leading humans to consider their decision more carefully.

In summary, we have created a POMDP model for estimating human trust and workload in the context of a human interacting with a decision-aid system. We observe that a higher transparency is not always the most likely way to increase trust in humans nor is it always more likely to increase workload. Instead, the optimal transparency depends on the current state of human trust and workload along with the recommendation type and the human's past experiences. In other words, higher transparency is not always beneficial, and instead, system transparency should be controlled based upon all these factors. In the next section, we use the POMDP model to develop an optimal control policy that varies system transparency to improve human-machine interaction performance objectives.

3.4 Controller Design

In the last section, we developed a partially observable Markov decision process (POMDP) framework for estimating human trust and workload as it changes with machine transparency. The model captures changes in trust and workload for contexts that involve interaction between a human and an intelligent decision-aid system. In this section, we establish a systematic method for shaping the reward function for the trust-workload POMDP model framework so as to close the loop between human and machine. We implement these control policies in a reconnaissance mission study in which human subjects are aided by a virtual robotic assistant. Finally, we analyze the performance of these two control policies against an open-loop baseline.

Before we synthesize an optimal control policy, we need to define the context-specific performance objectives relevant in this study. We focus on two critical per-

formance objectives: 1) the human should make correct decisions irrespective of the robot's reliability and 2) the human should make their decision in the shortest amount of time. Based on these performance objectives, we define the reward function for the POMDP, which is used to obtain the optimal control policy.

3.4.1 Decision Reward Function

The primary goal of calibrating trust and workload during human-machine interactions is to achieve the goals that are specific to the interaction. Since any given decision-aid system is never completely reliable, it is not always beneficial for the human to comply with the system. Instead, the human should make *correct* decisions; that is, the human should comply with the system when its recommendation is reliable and not comply when the system's recommendation is faulty. The decision reward function aims to enforce this behavior by appropriately penalizing the human's decisions. In order to characterize this behavior formally, we first define a few terms.

Earlier we defined the recommendation $a_{S_A} \in S_A := \{S_A^-, S_A^+\}$ of the decision-aid system depending on its inference about a given situation. However, we also need to distinguish the human's inference about the situation from the true situation. For example, in our reconnaissance mission, it is possible for the decision-aid robot to correctly recommend the use of Light Armor, indicating the absence of gunmen, i.e., S_A^-, but for the human to believe that the robot is faulty. In this case, the human may infer that there are gunmen present and choose to wear Heavy Armor. To account for this situation, we additionally define the true absence or presence of the stimulus as $\bar{a}_S \in S := \{S^-, S^+\}$ and the human's inference as $\bar{a}_{S_H} \in S_H := \{S_H^-, S_H^+\}$. Note that terms with $\bar{a}$ should not be confused with actions a of the POMDP model. Also, $\bullet^-$ and $\bullet^+$ represent the absence and presence of a stimulus, respectively. Typically, the prior probability of the true situation $p(\bar{a}_S)$ is known. Moreover, the presence or absence of gunmen is equally probable in our study, so $p(S^-) = p(S^+) = 0.5$.

Table 3.3.
Confusion matrix representation for the decision-aid system's and the human's inference. Each row of the matrix represents the true situation, while each column represents the inference made by the decision-aid system or the human.

		Decision-aid system's or human's inference	
		$a_{S_A} = S_A^-$ or $\bar{a}_{S_H} = S_H^-$	$a_{S_A} = S_A^+$ or $\bar{a}_{S_H} = S_H^+$
True Situation	$\bar{a}_S = S^-$	True Negative TN	False Positive FP
	$\bar{a}_S = S^+$	False Negative FN	True Positive TP

Table 3.4.
Reliability characteristics of the decision-aid system in the reconnaissance mission study representing the probabilities of the decision-aid's inference given the true situation. Since the decision-aid is 70% reliable, the probability of the decision-aid making a correct inference is 0.7.

		Decision-aid robot's inference	
		$a_{S_A} = S_A^-$	$a_{S_A} = S_A^+$
True Situation	$\bar{a}_S = S^-$	$1 - \alpha$ $= 0.7$	α $= 0.3$
	$\bar{a}_S = S^+$	β $= 0.3$	$1 - \beta$ $= 0.7$

The decision-aid's recommendations, and the human's decisions with respect to the true situation, are each characterized by the confusion matrix shown in Table 3.3. In practice, a decision-aid system's reliability is a system characteristic and known *a priori*; therefore, we define the reliability function as a probability of the system's recommendation given the true situation, i.e., $p(a_{S_A}|\bar{a}_S)$; we denote the probability of the decision-aid system making a false negative as $p(S_A^-|S^+) = \beta$ and the probability of the decision-aid system making a false positive as $p(S_A^+|S^-) = \alpha$. These reliability characteristics of the decision-aid system with 70% reliability in our reconnaissance mission study are summarized in Table 3.4. To help the human make correct decisions, we define a decision reward function $\mathcal{R}_D : S_H \times S \to \mathbb{R}$ in terms of the human inference

Table 3.5.
Decision reward function based on the inference made by the human. The reward function is defined as penalties equivalent to the expected amount of time, in seconds, that the human has to expend as a result of their decision.

		Human's inference	
		$\bar{a}_{S_H} = S_H^-$	$\bar{a}_{S_H} = S_H^+$
True Situation	$\bar{a}_S = S^-$	-3	-7
	$\bar{a}_S = S^+$	-23	-7

and the true situation, which is summarized in Table 3.5. The reward function is defined in terms of penalties equivalent to the expected amount of time, in seconds, that the human has to expend as a result of their decision. In particular, the human has to wait 3 seconds to search the building with Light Armor S_H^- if there are no gunmen present S^-. However, if gunmen are present S^+ and the human chooses Light Armor S_H^-, an additional 20 second penalty due to injury will be applied, resulting in a total wait time of 23 seconds. Moreover, a choice of Heavy Armor S_H^+ will always result in a wait of 7 seconds to search the building irrespective of the true situation. This reward function is specific to the reconnaissance study context and should be, in general, defined based on the context under consideration.

Although, the decision reward function $\mathcal{R}_D(\bar{a}_{S_H}, \bar{a}_S)$ is intuitive to design, the standard form of the reward function for a POMDP $\mathcal{R} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is defined as the reward for transitioning from state $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ due to action $a \in \mathcal{A}$. Therefore, we transform $\mathcal{R}_D(\bar{a}_{S_H}, \bar{a}_S)$ to derive the expected standard reward function. As decision rewards are only dependent on human compliance behavior and therefore only on trust behavior, we derive the expected reward function for the trust POMDP model as $\mathcal{R}_T : T \times T \times \mathcal{A} \to \mathbb{R}$ by calculating $\mathbf{E}[R|s_T, s'_T, a]$, where random variable R is the reward and $\mathbf{E}[\bullet]$ is the expected value of $\bullet$.

Proposition 3 *Given the reward function $\mathcal{R}_D(\bar{a}_{S_H}, \bar{a}_S)$ as shown in Table 3.5, trust emission probability function $\mathcal{E}_T(o_C, s_T)$, and automation reliability characteristics*

defined as $p(a_{S_A}|\bar{a}_S)$*, an equivalent expected reward function in the form* $\mathcal{R}_T(s_T, s'_T, a)$ *is*

$$\mathcal{R}_T(s_T, s'_T, [a_{S_A}, a_E, a_\tau]) = \sum_{o'_C \in C} \sum_{\bar{a}_S \in S} \mathcal{E}_T\left(o'_C|s'_T\right) p\left(\bar{a}_S|a_{S_A}\right) \mathcal{R}_D\left(g\left(a_{S_A}, o'_C\right), \bar{a}_S\right) \quad , \tag{3.1}$$

where $p\left(\bar{a}_S|a_{S_A}\right)$ *is the posterior probability calculated from Table 3.4 and Bayes' theorem as*

$$p\left(\bar{a}_S|a_{S_A}\right) = \frac{p\left(a_{S_A}|\bar{a}_S\right) p\left(\bar{a}_S\right)}{\sum\limits_{\bar{a}_S \in S} p\left(a_{S_A}|\bar{a}_S\right) p\left(\bar{a}_S\right)} \quad ,$$

and $g : S_A \times C \to S_H$ *is a function mapping human compliance* $o_C \in C$ *in response to the system's recommendation* $a_{S_A} \in S_A$ *to the human inference/decision* $\bar{a}_{S_H} \in S_H$*. Specifically,*

$$g\left(S_A^-, C^-\right) = S_H^+ \ , \ g\left(S_A^-, C^+\right) = S_H^- \ , \ g\left(S_A^+, C^-\right) = S_H^- \ , \ \textit{and} \ g\left(S_A^+, C^+\right) = S_H^+ \ .$$

For example, the human not complying $o_C = C^-$ *with a recommendation of Light Armor* $a_{S_A} = S_A^-$ *effectively means that the human is inferring the presence of gunmen, and thereby, choosing Heavy Armor* $\bar{a}_{S_H} = S_H^+$*.*

Proof Let the reward R be a random variable, which has been defined in terms of human inference $\bar{a}_{S_H} \in S_H$ and the true situation $\bar{a}_S \in S$ (see Table 3.5). We derive the expected reward function for the trust POMDP model as $\mathcal{R}_T : T \times T \times \mathcal{A} \to \mathbb{R}$ by calculating $\mathbf{E}\left[R|s_T, s'_T, a\right]$, where $\mathbf{E}\left[\bullet\right]$ is the expected value of $\bullet$. Therefore, using

the law of total expectation repeatedly, as well as conditional independence between states, actions, and observations, we obtain

$$\begin{aligned}
\mathcal{R}_T(s_T, s'_T, [a_{S_A}, a_E, a_\tau]) &= \mathbf{E}\left[R|s_T, s'_T, a_{S_A}, a_E, a_\tau\right] \\
&= \mathbf{E}\left[R|s'_T, a_{S_A}\right] \\
&= \sum_{o'_C \in C} p\left(o'_C|s'_T, a_{S_A}\right) \mathbf{E}\left[R|s'_T, a_{S_A}, o'_C\right] \\
&= \sum_{o'_C \in C} p\left(o'_C|s'_T\right) \mathbf{E}\left[R|a_{S_A}, o'_C\right] \\
&= \sum_{o'_C \in C} \sum_{\bar{a}_S \in S} \mathcal{E}_T\left(o'_C|s'_T\right) p\left(\bar{a}_S|a_{S_A}, o'_C\right) \mathbf{E}\left[R|a_{S_A}, o'_C, \bar{a}_S\right] \\
&= \sum_{o'_C \in C} \sum_{\bar{a}_S \in S} \mathcal{E}_T\left(o'_C|s'_T\right) p\left(\bar{a}_S|a_{S_A}\right) \mathcal{R}_D\left(g\left(a_{S_A}, o'_C\right), \bar{a}_S\right) \quad .
\end{aligned}$$

■

3.4.2 Response Time Reward Function

In most scenarios, apart from ensuring that the human makes correct decisions, the time the human takes to make the decision is also critical. Therefore, to minimize the human's response time, we define the response time reward function $\mathcal{R}_{RT} : \mathbb{R}^+ \to \mathbb{R}$ as $\mathcal{R}_{RT}\left(o_{RT}\right) = -o_{RT}$ to proportionally penalize longer response times. Similar to the decision reward function, the response time reward function is transformed to derive the expected standard reward function for the POMDP. As the response time reward function is only dependent on human response time behavior, and therefore, only on workload behavior, we derive the expected reward function for the workload POMDP model as $\mathcal{R}_W : W \times W \times \mathcal{A} \to \mathbb{R}$ by calculating $\mathbf{E}\left[R|s_W, s'_W, a\right]$, where random variable R is the reward and $\mathbf{E}\left[\bullet\right]$ is the expected value of $\bullet$.

Proposition 4 *Given the reward function* $\mathcal{R}_{RT}(o_{RT}) = -o_{RT}$ *and workload emission probability function* $\mathcal{E}_W(o_{RT}|s_W)$ *as represented in Figure 3.7, an equivalent expected reward function in the form* $\mathcal{R}_W(s_W, s'_W, a)$ *is*

$$\mathcal{R}_W(s_W, s'_W, [a_{S_A}, a_E, a_\tau]) = -(\mu_{s'_W} + \tau_{s'_W}), \tag{3.2}$$

where $\mu_{s'_W}$ *and* $\tau_{s'_W}$ *are the parameters of the ex-Gaussian distribution corresponding to state* $s'_W \in W$.

Proof Let the reward R be a random variable, which has been defined in terms of human response time $o_{RT} \in \mathbb{R}^+$. We derive the expected reward function for the workload POMDP model as $\mathcal{R}_W : W \times W \times \mathcal{A} \to \mathbb{R}$ by calculating $\mathbf{E}[R|s_W, s'_W, a]$, where $\mathbf{E}[\bullet]$ is the expected value of $\bullet$. Therefore, using the law of total expectation repeatedly, as well as conditional independence between states, actions, and observations, we obtain

$$\begin{aligned}
\mathcal{R}_W(s_W, s'_W, [a_{S_A}, a_E, a_\tau]) &= \mathbf{E}[R|s_W, s'_W, a_{S_A}, a_E, a_\tau] \\
&= \int_{\mathbb{R}^+} p(o'_{RT}|s_W, s'_W, a_{S_A}, a_E, a_\tau)\, \mathbf{E}[R|o'_{RT}, s_W, s'_W, a_{S_A}, a_E, a_\tau]\, do'_{RT} \\
&= \int_{\mathbb{R}^+} p(o'_{RT}|s'_W)\, \mathbf{E}[R|o'_{RT}]\, do'_{RT} \\
&= \int_{\mathbb{R}^+} p(o'_{RT}|s'_W)\, (-o_{RT})\, do'_{RT} \\
&= -\mathbf{E}[o'_{RT}|s'_W] \\
&= -(\mu_{s'_W} + \tau_{s'_W}) \ .
\end{aligned}$$

■

We define the total reward function $\mathcal{R}$ for human trust-workload behavior as a convex combination of (3.1) and (3.2) with weight ζ as

$$\mathcal{R} = \zeta \mathcal{R}_T + (1 - \zeta)\mathcal{R}_W \ . \tag{3.3}$$

As the weight ζ increases, more importance is given to the trust reward than the workload reward. For situations in which a correct decision is more important than a faster response time, a higher value of ζ should be used. Lastly, the discount factor γ is selected based on the number of trials per mission in our study, i.e., $N = 15$. We select the discount factor γ such that the reward of the 15^{th} trial has a weight of e^{-1}; such a value of γ can be approximated as

$$\gamma = \frac{N}{N+1} = 0.9375 \ . \tag{3.4}$$

With the defined reward function and discount factor, we calculate the control policy for the POMDP model using the Q-MDP method as described in the next section.

3.4.3 POMDP Control Policy

Using the reward function defined in the previous section, we determine the optimal control policy for updating the decision-aid's transparency by solving the combined trust-workload model to maximize the reward function defined in the previous section. Although it is possible to obtain the exact solution of the optimization through dynamic programming using value iteration, the time complexity increases exponentially with the cardinality of the action and observation spaces. Since a real-world scenario can involve a much larger set of actions and observations, obtaining the exact optimal solution may be intractable. Therefore, we adopt an approximate greedy approach called the Q-MDP method [188] to obtain a near optimal transparency control policy. The Q-MDP method solves the underlying MDP by ignoring the observation probability function to obtain the Q-function $Q_{\text{MDP}} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$.

$Q_{\text{MDP}}(s,a)$ is the optimal expected reward given an action a is taken at the current state s. Then, using the belief state $b(s)$, which can be iteratively calculated as

$$b'(s') = p(s'|o,a,b(s)) = \frac{p(o|s',a)\sum_{s\in\mathcal{S}} p(s'|s,a)b(s)}{\sum_{s'\in\mathcal{S}} p(o|s',a)\sum_{s\in\mathcal{S}} p(s'|s,a)b(s)} \ , \tag{3.5}$$

the optimal action a^* is chosen as

$$a^* = \underset{a}{\text{argmax}} \sum_{s\in\mathcal{S}} b(s)Q_{\text{MDP}}(s,a) \ . \tag{3.6}$$

Essentially, the Q-MDP method approximates the optimal solution by assuming that the POMDP becomes completely observable after the next action. In order to solve the POMDP using the Q-MDP method, we calculate the Q-function of the underlying MDP using value iteration [169]. Nevertheless, as with any other method, the solution assumes that the decision-aid system can take any action $a \in \mathcal{A}$ in the future. But, in our model, only transparency a_τ is a controllable action; the other actions—recommendation a_{S_A} and experience a_E—depend on the context and cannot be explicitly controlled by the policy. To account for these "uncontrollable" actions while solving for the control policy in the Q-MDP method, we calculate an expected Q-function of the form $Q^\tau : \mathcal{S} \times \tau \to \mathbb{R}$. This intermediate Q-function is only dependent on the controllable actions and considers the probabilities of the uncontrollable actions. Finally, we iteratively solve (3.7) until convergence is achieved to obtain $Q_{\text{MDP}}(s,a)$.

$$\begin{aligned} Q_{\text{MDP}}(s,a) &= \sum_{s'\in\mathcal{S}} \mathcal{T}(s'|s,a)\left(\mathcal{R}(s'|s,a) + \gamma V(s')\right) \\ Q^\tau(s,\tau) &= \sum_{a_{S_A}\in S_A, a_E\in E} p(a_{S_A},a_E)Q_{\text{MDP}}(s,[a_{S_A},a_E,a_\tau]) \\ V(s) &= \max_\tau Q^\tau(s,\tau) \end{aligned} \tag{3.7}$$

Furthermore, $p(a_{S_A}, a_E) = p(a_{S_A})p(a_E)$ because the present recommendation a_{S_A} and experience a_E due to the reliability of the last recommendation are independent. Therefore, $p(a_{S_A})$ and $p(a_E)$ are calculated as

$$\begin{aligned} p({S_A}^-) &= \beta d + (1-\alpha)(1-d) \ , \\ p({S_A}^+) &= 1 - p(S_A^-) \ , \\ p(E^-) &= \alpha(1-d) + \beta d \ , \\ p(E^+) &= 1 - p(E^-) \ . \end{aligned} \tag{3.8}$$

For our human subject study, $d = 0.5$, $\alpha = 0.3$, and $\beta = 0.3$. For implementation, once a_{S_A} and a_E are known in a trial, near-optimal transparency a_τ^* can be determined as

$$a_\tau^* = \underset{a_\tau}{\operatorname{argmax}} \sum_{s \in \mathcal{S}} b(s) Q_{\mathrm{MDP}}(s, [a_{S_A}, a_E, a_\tau]) \ . \tag{3.9}$$

We calculate the total reward function $\mathcal{R}$ and the corresponding control policy for three values of reward weights $\zeta = 0.50$, $\zeta = 0.91$, and $\zeta = 0.95$.

The control policies corresponding to each of the reward weights are depicted in Figures 3.9, 3.10, and 3.11, respectively. We first consider the case with $\zeta = 0.50$ shown in Figure 3.9. Here, the reward function gives equal importance to the decision and response time rewards. Each of the four figures represents the optimal choice of transparency based on the estimated probability of High Trust $T_\uparrow$ and High Workload $W_\uparrow$ for a given recommendation a_{S_A} and experience a_E. We first consider the case when the recommendation suggests Light Armor $a_{S_A} = S_A^-$ as shown in Figures 3.9(a) and 3.9(b). This case represents a high risk situation for over-trust because an incorrect human decision of complying with the recommendation can lead to the human using Light Armor in the presence of gunmen, resulting in injury and an extra penalty of 20 seconds. The control policy adopts medium transparency τ_M when the probabilities of High Trust and High Workload are high. Medium transparency can help the human to make a more informed decision than low transparency, thereby

avoiding over-trust in these cases when the human's trust is too high. Also, if the human's experience was reliable from the last trial ($a_E = E^+$), the chance of over-trust is higher. In this case, medium transparency is adopted at even lower probabilities of High Trust as seen in Figure 3.9(b). Furthermore, as seen in Figure 3.8, medium transparency is best at transitioning a human from a state of High Workload to a state of Low Workload. Therefore, medium transparency will help to reduce the expected response time when the probability of High Workload is high.

For the case when the decision-aid recommends Heavy Armor $a_{S_A} = S_A^+$, shown in Figures 3.9(c) and 3.9(d), the situation risk is low given that an incorrect compliance only leads to an extra 4 seconds of search time. Therefore, the control policy always aims to increase trust in this case. Since medium transparency has the highest probability of causing a transition from Low Trust to High Trust (Figure 3.6(c)) and 3.6(d)), the control policy adopts medium transparency when the probability of High Trust is low. When the probability of High Trust is high and the human's prior experience with the decision aid was Faulty, medium transparency has a higher probability of maintaining a high trust level as compared to low transparency (Figure 3.6(c)); therefore, the control policy adopts medium transparency in this case (Figures 3.9(c)). Note that high transparency is not adopted by the control policy in this case due to the large response time penalty associated with high transparency. When the human's prior experience with the decision aid was reliable, low transparency has the highest probability of maintaining high trust level (Figure 3.6(d)); therefore, low transparency is preferred with low levels of workload (Figures 3.9(d)). Moreover, medium transparency is adopted when the probability of High Workload is high as discussed above. In general, medium transparency dominates the control policy for $\zeta = 0.50$ because in most cases it provides a good trade-off between trust calibration based on informed decision-making and increased workload.

For the cases with $\zeta = 0.91$ and $\zeta = 0.95$, higher importance is given to the decision rewards as compared to the response time rewards. In these cases, as represented in Figure 3.10 and 3.11, we observe that the control policies adopt high transparency for

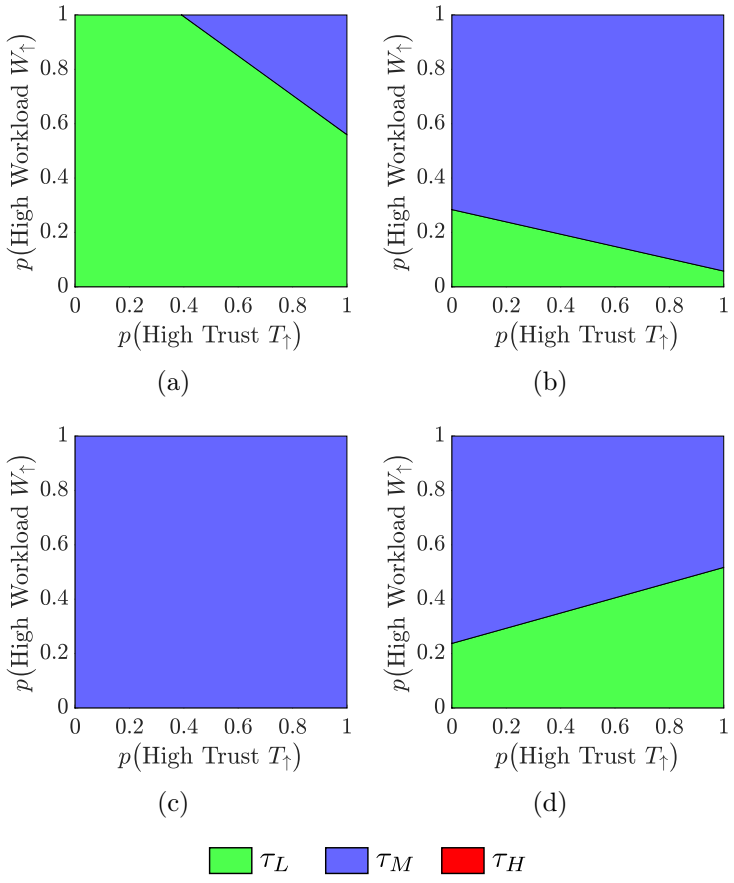


Figure 3.9. Closed-loop control policy corresponding to the reward function with $\zeta = 0.50$. In this case, the reward function gives equal importance to the decision and response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$. When $\zeta = 0.50$, high transparency is never adopted because it would result in a significant increase in response time.

a very high probability of High Trust; this ensures that the human does not over-trust the automation and instead makes the most informed decision possible. With higher values of ζ, the use of high transparency is further increased since the associated weight for the response time reward is significantly reduced. In the next section, these control policies are implemented to dynamically vary transparency based on the participant's current trust and workload estimates in a reconnaissance mission study.

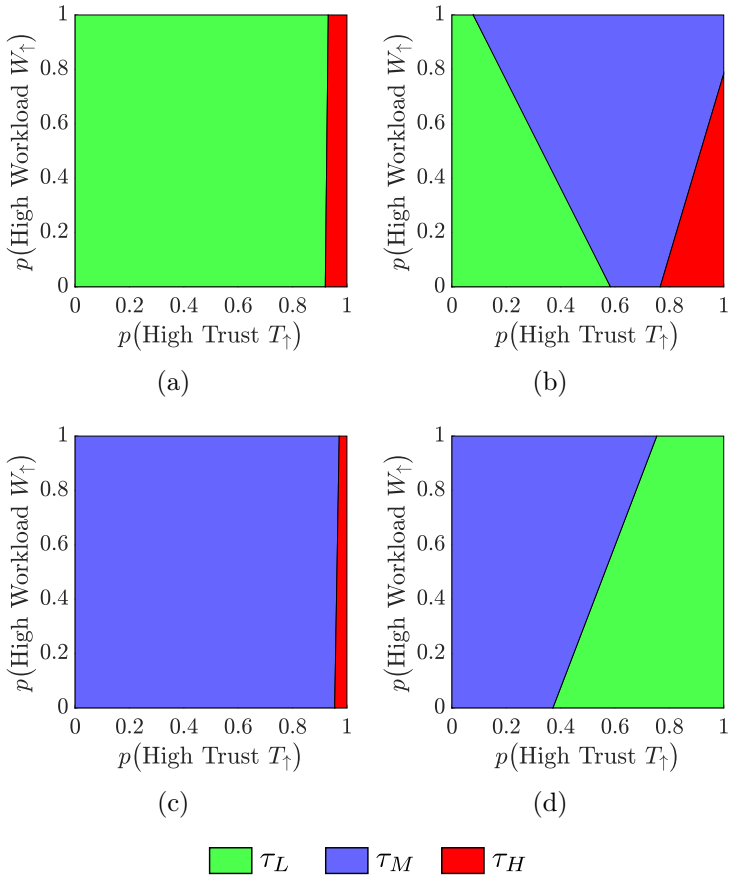


Figure 3.10. The closed-loop control policy corresponding to the reward function with $\zeta = 0.91$. In this case, higher importance is given to the decision rewards as compared to the response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$. This control policy adopts high transparency for very high probabilities of High Trust to reduce the number of incorrect decisions the human may make due to their over-trust in the decision-aid system.

3.5 Validation and Results

To experimentally validate the performance of the proposed control policies represented in Figures 3.9, 3.10, and 3.11, we conducted two human subject studies. These experiments were identical to the one used to collect open-loop data for each

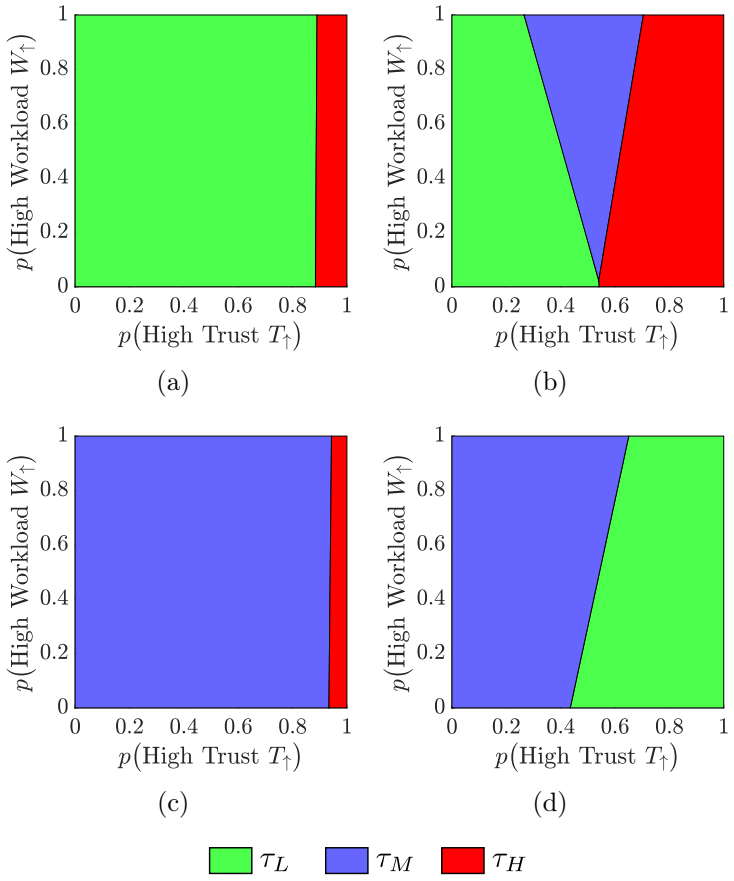


Figure 3.11. The closed-loop control policy corresponding to the reward function with $\zeta = 0.95$. In this case, a very high importance is given to the decision rewards as compared to the response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$. This control policy again adopts high transparency for high probabilities of High Trust to reduce the number of incorrect decisions the human may make due to their over-trust in the decision-aid.

transparency but with transparency controlled using the control policies based on human trust and workload estimation.

Stimuli and Procedure: Two within-subject studies were performed in which participants were asked to interact with a simulation of three reconnaissance missions as described in the earlier study description. However, instead of fixed transparency

in each mission, the transparency was controlled based on a feedback control policy for some missions. For the first study, high transparency was always used in one of the three missions, and in the other two missions, the transparency was dynamically varied based on control policies corresponding to $\zeta = 0.50$ (Figure 3.9) and $\zeta = 0.95$ (Figure 3.11), respectively. In the second study, the transparency was dynamically varied based on the control policy corresponding to $\zeta = 0.91$ (Figure 3.10) in one mission, and the other two missions used fixed medium transparency and fixed high transparency, respectively. Three missions in each study ensured that the studies were short enough to avoid participant fatigue and were consistent in structure with the study used to collect open-loop data. Moreover, the order of missions was again randomized across participants to reduce ordering effects [185].

Participants: One hundred and twenty participants for the first study, and one hundred and four participants for the second study, participated in and completed the study online. They were recruited using Amazon Mechanical Turk [97] with the same criteria used for the earlier study. To account for participants who were not sufficiently engaged in the study, we filtered data that had any response time higher than 40.45 seconds. As a result, 20 outlying participants were removed from the dataset for the first study, and 7 outlying participants were removed from the dataset for the second study, leading to a remaining 100 and 97 participants, respectively.

Using the collected human subject data from the two validation studies along with the open-loop study discussed earlier, we quantify and evaluate participants' performance for the dynamically varying transparency missions and the fixed transparency mission. We compare two metrics: total decision reward and total response time reward for each type of mission. We use linear mixed effects analysis and likelihood ratio tests to determine whether the use of trust-workload behavior-based feedback had any significant effect on these metrics. We used the statistical computing language R [189] and *lme4* library [190] to perform a linear mixed model approach to analyze the relationship between each of the metrics and the transparency policies. As a fixed effect, we used the transparency policy in the models. To account for

variations in the metrics calculated for different participants, the models considered each individual as a random effect. P-values were obtained using likelihood ratio tests of the full model that includes the transparency policy as a fixed effect against the model that does not include the transparency policy. Figure 3.12 shows the effect of the transparency policies (open-loop: Low, Medium, and High; closed-loop: $\zeta = 0.50, 0.91,$ and 0.95) on the total decision rewards and on the total response time rewards across participants.

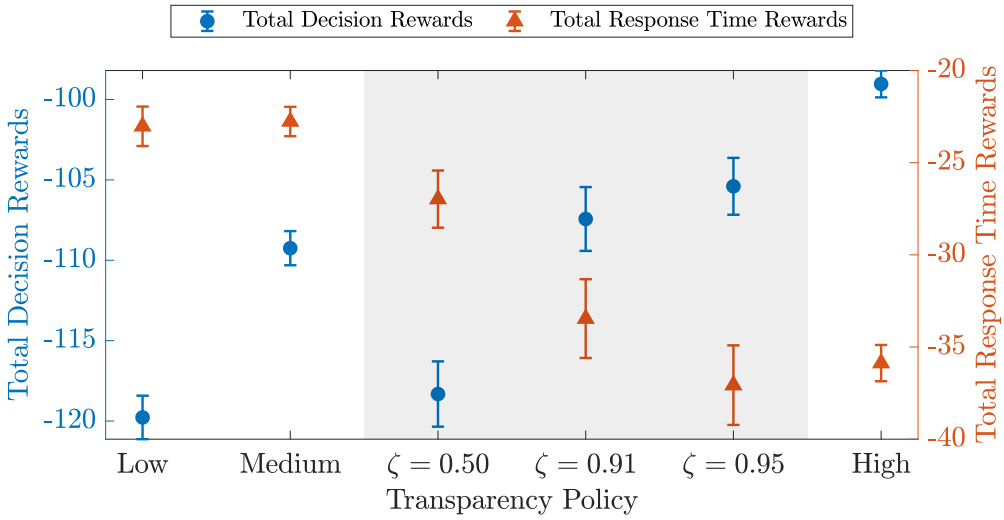


Figure 3.12. Effect of the proposed control policies on the total decision and total response time rewards. Error bars represent the standard error of the mean across participants. The closed-loop control policies are highlighted in gray. The performance of the closed-loop policies lies between that of high and low transparency in terms of both reward metrics. With higher values of the reward weight ζ, the performance of the closed-loop policy is more similar to that of high transparency. Depending on the requirements of the context, ζ can be tuned to achieve the required trade-off between decision and response time performance.

The total decision reward is defined as the sum of all decision rewards based on Table 3.5 accrued by the participant in a mission. A likelihood ratio test using linear mixed effects models indicated that the transparency policies significantly affected total decision rewards ($\chi^2(6) = 229.62, p \approx 0.0000$). The total response time reward is

defined as the negative of the sum of all response times in seconds accrued by the participant in a mission. A likelihood ratio test indicated that the transparency policies significantly affected total response time rewards ($\chi^2(6) = 230.07, p \approx 0.0000$).

To analyze the performance and benefit of the closed-loop control policies, we compare them with the performance of the open-loop cases (that consider a static transparency). Furthermore, we analyze the effects of the reward weight ζ on the closed-loop performance. From Figure 3.12, considering open-loop fixed transparency policies, we see that high transparency has the best performance in terms of decision rewards, followed by medium and low transparency. However, low and medium transparency perform better in terms of response time rewards; this indicates a trade-off between the correctness of the human's decision versus the corresponding response time. Although the use of high transparency can result in the highest number of correct decisions, high response times indicate higher workload levels for the human. Furthermore, some time-critical contexts may favor fast response times in lieu of perfect decisions.

Given this trade-off, we see that the performance of our closed-loop policies lies between that of high and low transparency in terms of both reward metrics (see gray-highlighted region in Figure 3.12). We see that with higher values of the weight ζ, the performance of the closed-loop policy is more similar to that of high transparency used all the time. The control policy corresponding to $\zeta = 0.91$ performs better than the medium transparency in terms of decision rewards but has lower response time rewards. Therefore, depending on the requirements of the context, the proposed control policy enables the controls engineer to trade off between decision and response time performance.

3.6 Chapter Summary

In this chapter, we developed a model of human trust and workload dynamics as they evolve during a human's interaction with a decision-aid system. Furthermore,

we designed and validated a model-based feedback control policy aimed at dynamically varying the automation's transparency to improve the overall performance of the human-machine team. The model, which was parameterized using human subject data, captured the effects of the decision-aid's recommendation, the human's previous experience with the automation, and automation transparency on the human's trust-workload behavior. The model is capable of estimating human trust and workload in real time using recursive belief-state estimates. Experimental validation showed that the closed-loop control policies were successfully able to manage the human decision versus response time performance tradeoff based on a tuning parameter in the reward function. This framework provides a tractable methodology for using human behavior as a real-time feedback signal to optimize human-machine interactions through dynamic modeling and control.

It should be noted that the overall performance of the control policy could be improved by addressing a few limitations of the proposed trust-workload model. We assumed that human trust and workload behavior are conditionally independent to simplify the model structure and complexity. However, trust and workload may be coupled, and therefore, changes in the trust state could directly impact the workload state and vice-versa. In the next chapter, we will explore and analyze coupled models of trust and workload by relaxing multiple independence assumptions.

4. COUPLED MODELS OF TRUST AND WORKLOAD

In the previous chapter, we demonstrated the use of a POMDP based framework to model the dynamics of human trust and workload. We then used this model to design a feedback policy to optimally vary automation transparency. As we noted, a potential limitation of this framework is that it assumes that human trust and workload dynamics are independent, an assumption that is accompanied by the notion that human compliance is not affected by workload and response time is not affected by trust. However, studies in existing work contradict some of these assumptions. First, researchers have shown that workload can have an impact on human trust behavior [69, 191], and that trust can facilitate reliance and reduce the workload associated with monitoring the automation [192]. Additionally, [66] notes that compliance affects the response time and accuracy to an announced system failure (higher compliance leads to shorter response time), both under high workload conditions. Therefore, we need to analyze the coupling interactions between human trust and workload in the context of human-machine interactions.

In this chapter, we model this coupling in a POMDP framework for a human interacting with an automated decision-aid. We specifically consider a simulated reconnaissance mission scenario where the human is assisted by a robot as discussed in Chapter 3. We explore and analyze multiple models with varying complexities by relaxing assumptions on trust and workload independence. Finally, the performance of two of these coupled models are compared to that of the independent model by validating the optimal control policy for dynamically varying automation transparency based on each model's trust and workload estimates.

4.1 Description of Coupled Models

The trust-workload POMDP model as defined in Chapter 3 is as follows. The set of states $\mathcal{S}$ is defined as tuples containing the *Trust* state s_T and the *Workload* state s_W, i.e., $s = \Big[s_T, s_W\Big]$. Here,

$$s_T \in T := \Big\{\text{Low Trust } T_{\downarrow}, \text{High Trust } T_{\uparrow}\Big\} \text{ and}$$
$$s_W \in W := \Big\{\text{Low Workload } W_{\downarrow}, \text{High Workload } W_{\uparrow}\Big\} \ .$$

The set of actions $\mathcal{A}$ is defined as the characteristics of the system recommendation that consists of tuples containing *Recommendation* of the automation a_{S_A}, *Experience* a_E that depends on the reliability of the last recommendation, and *Transparency* a_τ, i.e., $a = \Big[a_{S_A}, a_E, a_\tau\Big]$. Here,

$$a_{S_A} \in S_A := \Big\{\text{Stimulus Absent } S_A^-, \text{Stimulus Present } S_A^+\Big\} \ ,$$
$$a_E \in E := \begin{Bmatrix}\text{Faulty last experience } E^-, \\ \text{Reliable last experience } E^+\end{Bmatrix} \text{, and}$$
$$a_\tau \in \tau := \begin{Bmatrix}\text{Low Transparency } \tau_L, \\ \text{Medium Transparency } \tau_M, \\ \text{High Transparency } \tau_H\end{Bmatrix} \ .$$

The observable characteristics of the human's decision are defined as a set of observations $\mathcal{O}$ consisting of tuples containing *Compliance* o_C and *Response Time* o_{RT}, i.e., $o = \Big[o_C, o_{RT}\Big]$. Here,

$$o_C \in C := \Big\{\text{Disagree } C^-, \text{Agree } C^+\Big\}$$

and $o_{RT} \in \mathbb{R}^+$ is defined as the time the human takes to respond after receiving the decision-aid's recommendation. The transition from the current state $s \in \mathcal{S}$ to the next state $s' \in \mathcal{S}$ given the action $a \in \mathcal{A}$ is characterized by the transition

probability function $\mathcal{T}(s'|s,a)$. The emission probability function $\mathcal{E}(o|s)$ characterizes the likelihood of observing $o \in \mathcal{O}$ given the process is in state s. The transition probability function and the emission probability function for our trust-workload model are given by

$$\mathcal{T}(s'|s,a) = \mathcal{T}(s'_T, s'_W|s_T, s_W, a) \text{ and}$$
$$\mathcal{E}(o|s) = \mathcal{E}(o_C, o_{RT}|s_T, s_W) \ .$$

We will explore four types of coupled models of trust and workload by relaxing the independence assumptions established in Chapter 3. These models are described in order of increasing number of parameters and thereby, increasing complexity.

4.1.1 Independent Model

The model we defined in Section 3.2 has the lowest complexity; hereafter, we refer to this model as the independent model. For the independent model, we assumed that the states of trust and workload are independent. We also assumed that the observations *compliance* and *response time* are only dependent on trust and workload, respectively. Therefore, the transition and emission probability functions can be simplified as:

$$\begin{aligned} \mathcal{T}(s'|s,a) &= \mathcal{T}(s'_T|s_T,a)\mathcal{T}(s'_W|s_W,a) \\ \mathcal{E}(o|s) &= \mathcal{E}(o_C|s_T)\mathcal{E}(o_{RT}|s_W) \end{aligned} \tag{4.1}$$

This model is represented in Figure 4.1. Here, the transition and observation probability functions consisting only of discrete variables, i.e., $\mathcal{T}(s'_T|s_T,a)$, $\mathcal{T}(s'_W|s_W,a)$, and $\mathcal{E}(o_C|s_T)$, are modeled as multinomial distributions. The observation probability function consisting of continuous response time, i.e., $\mathcal{E}(o_{RT}|s_W)$, is modeled as a set of two (one for each state of workload) exponentially modified Gaussian distributions as described in Chapter 3. For the multinomial distributions, due to the constraint that the probabilities sum to one for a given starting state (and action for the transi-

tion probability function), the number of independent parameters is smaller than the absolute number of parameters. Considering this constraint, the effective number of parameters in this model is 58.

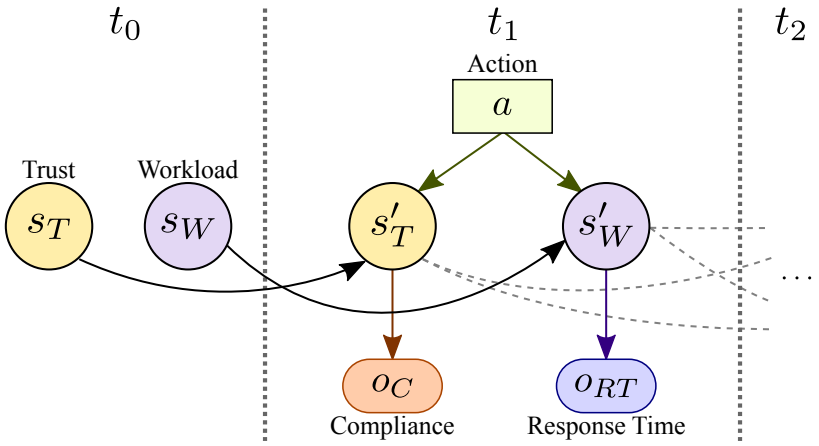


Figure 4.1. A representation of the independent model of trust and workload. The observations compliance and response time are only dependent on trust and workload, respectively.

4.1.2 Coupled-Transition Model

In this model, we relax one of the independence assumptions and allow the transition of trust and workload to be dependent on both of the previous states of trust and workload. Therefore, the transition and emission probability functions can be represented as:

$$\begin{aligned} \mathcal{T}(s'|s,a) &= \mathcal{T}(s'_T|s_T,s_W,a)\mathcal{T}(s'_W|s_T,s_W,a) \\ \mathcal{E}(o|s) &= \mathcal{E}(o_C|s_T)\mathcal{E}(o_{RT}|s_W) \end{aligned} \tag{4.2}$$

This model is represented in Figure 4.2. Again, $\mathcal{T}(s'_T|s_T,s_W,a)$, $\mathcal{T}(s'_W|s_T,s_W,a)$, and $\mathcal{E}(o_C|s_T)$ are modeled as multinomial distributions, and $\mathcal{E}(o_{RT}|s_W)$ is modeled as a set

of two (one for each state of workload) exponentially modified Gaussian distributions. The effective number of parameters in this model is 106.

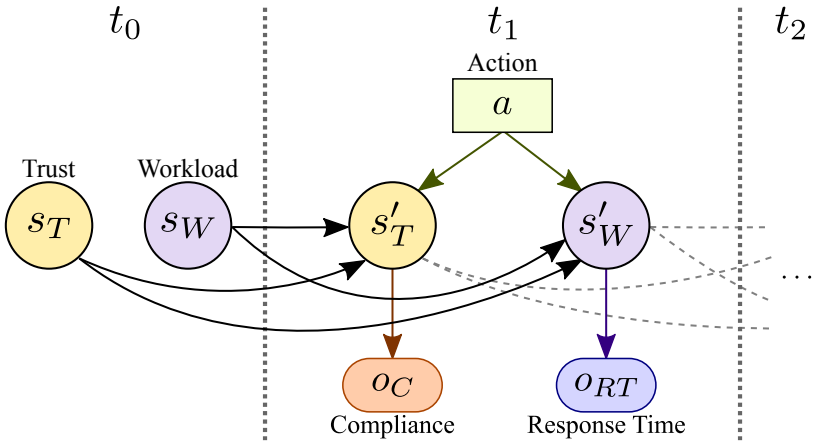


Figure 4.2. A representation of the coupled-transition model of trust and workload. The transition probabilities of trust and workload are dependent on both of the previous states of trust and workload.

4.1.3 Coupled-Emission Model

In this model, we relax another independence assumption and allow the emission probability functions of compliance and response time to be dependent on both the trust and workload states. Therefore, the transition and emission probability functions can be represented as:

$$\begin{aligned} \mathcal{T}(s'|s,a) &= \mathcal{T}(s'_T|s_T,s_W,a)\mathcal{T}(s'_W|s_T,s_W,a) \\ \mathcal{E}(o|s) &= \mathcal{E}(o_C|s_T,s_W)\mathcal{E}(o_{RT}|s_T,s_W) \end{aligned} \tag{4.3}$$

The model is represented in Figure 4.3. Again, $\mathcal{T}(s'_T|s_T,s_W,a)$, $\mathcal{T}(s'_W|s_T,s_W,a)$, and $\mathcal{E}(o_C|s_T,s_W)$ are modeled as multinomial distributions, and $\mathcal{E}(o_{RT}|s_T,s_W)$ is modeled as a set of four (one for each combination of trust and workload state) exponentially

modified Gaussian distributions. The effective number of parameters in this model is 114.

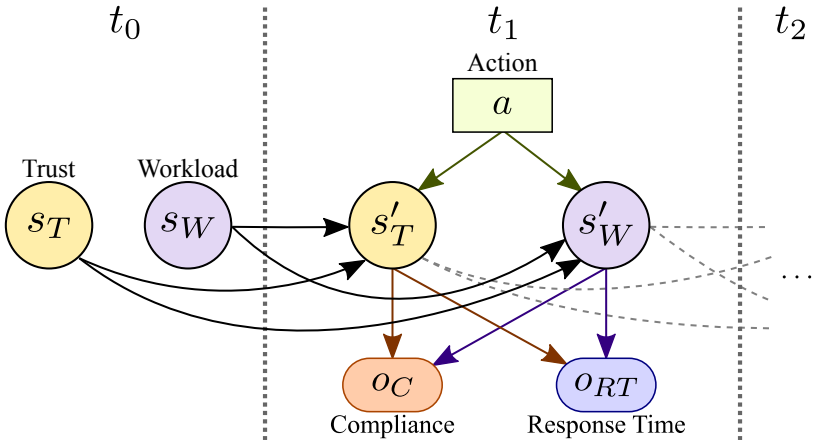


Figure 4.3. A representation of the coupled-emission model of trust and workload. The emission probability functions of compliance and response time are dependent on both the trust and workload states.

4.1.4 Coupled-State Model

In this model, we further relax the assumption that the states of trust and workload, at any given time, are independent. Therefore, the transition and emission probability functions can be represented as:

$$\begin{aligned} \mathcal{T}(s'|s,a) &= \mathcal{T}(s'_T, s'_W|s_T, s_W, a) \\ \mathcal{E}(o|s) &= \mathcal{E}(o_C|s_T, s_W)\mathcal{E}(o_{RT}|s_T, s_W) \end{aligned} \tag{4.4}$$

This model is represented in Figure 4.4. Again, $\mathcal{T}(s'_T, s'_W|s_T, s_W, a)$ and $\mathcal{E}(o_C|s_T, s_W)$ are modeled as multinomial distributions, and $\mathcal{E}(o_{RT}|s_T, s_W)$ is modeled as a set of four (one for each combination of trust and workload state) exponentially modified Gaussian distributions. The effective number of parameters in this model is 163.

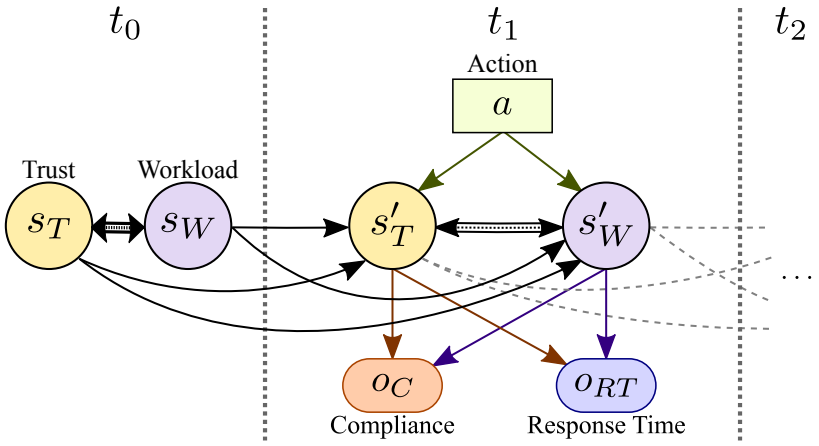


Figure 4.4. A representation of the coupled-state model for trust and workload.

4.1.5 Complete-Coupled Model

In this model, we assume that both states (trust and workload) are coupled and that both observations (compliance and response time) are also coupled. Therefore, the transition and emission probability functions can be represented as:

$$\begin{aligned} \mathcal{T}(s'|s,a) &= \mathcal{T}(s'_T, s'_W|s_T, s_W, a) \\ \mathcal{E}(o|s) &= \mathcal{E}(o_C, o_{RT}|s_T, s_W) \\ &= \mathcal{E}(o_C|s_T, s_W)\mathcal{E}(o_{RT}|o_C, s_T, s_W) \end{aligned} \tag{4.5}$$

This model is represented in Figure 4.5. Again, $\mathcal{T}(s'_T, s'_W|s_T, s_W, a)$ and $\mathcal{E}(o_C|s_T, s_W)$ are modeled as multinomial distributions, and $\mathcal{E}(o_{RT}|o_C, s_T, s_W)$ is modeled as a set of eight (one for each combination of compliance, trust, and workload) exponentially modified Gaussian distributions. The effective number of parameters in this model is 175.

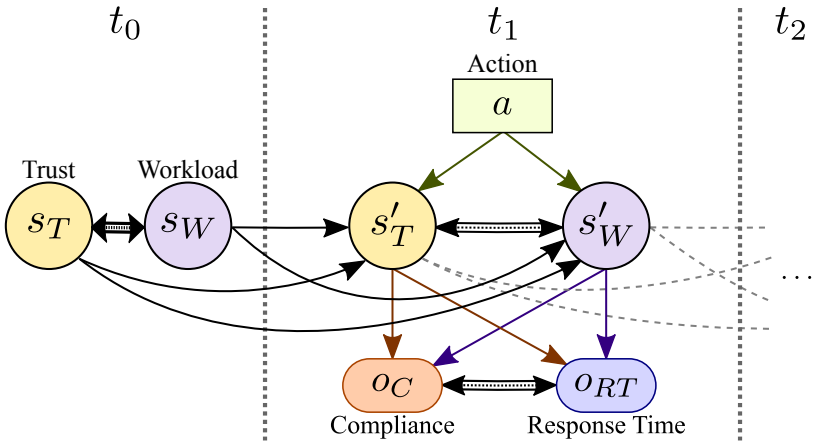


Figure 4.5. A representation of the complete-coupled model of trust and workload. No independence assumptions are made in this model.

4.2 Model Selection

Among the five trust-workload models described above, we now identify the models that best capture the trust-workload behavior exhibited by the participants in the collected data. We use the open-loop data collected in the reconnaissance mission study as described in Section 3.2.2. The data consists of 196×3 sequences of input-output data (196 participants with 3 missions each) where each sequence has a length of 15 trials. Each sequence is the action-observation data for a participant in the interaction between the participant and the robot in a given mission.

To obtain a model with the best generalizability given the available data, we calculate the five-fold cross validation log-likelihood for each model. Five-fold cross validation is done as follows. The data is randomly divided into five equal sets (called folds). This division is done by randomly creating five subsets out of the 196 participants; each fold comprises the data from one subset. For each fold, the model parameters are estimated using the aggregated data from the rest of the folds, and the log-likelihood of observing the sequences in the fold given the estimated model is calculated. The cross-validation log-likelihood is defined as the average of the log-likelihoods across the five folds. Furthermore, in order to increase the robustness of

the obtained validation log-likelihood values to variations in training and testing data sets, we repeat this process ten times to obtain the average five-fold cross validation log-likelihood for each model.

We use the genetic algorithm with the forward algorithm, using the methodology described in Section 3.3, to estimate the model parameters and obtain the average five-fold cross-validation log-likelihood for the five models. The results from ten iterations are shown in Figure 4.6. Two-sample t-tests show that the log-likelihood of the independent model is not significantly different from that of the coupled-transition model ($t(98) = 0.4245, p = 0.6722$) nor the coupled-emission model ($t(98) = -0.5975, p = 0.5516$). Moreover, the log-likelihood of the independent model is significantly *better* than that of coupled-state model ($t(98) = 3.9470, p = 0.0001$) and complete-coupled model ($t(98) = 7.4955, p \approx 0.0000$). This suggests that the coupled-state and complete-coupled models are too complex and are actually overfit to the training data, leading to a lower cross-validation accuracy.

Based on this cross-validation analysis of our data, we can conclude that at a given time sample, it is reasonable to assume that

- the human trust and workload states are conditionally independent given the previous trust-workload states and actions, and
- human compliance and response time are conditionally independent given the trust and workload states.

4.3 Model Parameter Estimation

Since the independent, coupled-transition, and coupled-emission models are not significantly different from each other (in terms of their cross-validation likelihoods based on open-loop human subject study data), we now evaluate their performance in a closed-loop human subject study. To do so, we conduct parameter estimation for the three models using the entire open-loop data with the methodology discussed in Section 3.3. Complete details about the estimated models are presented in Appendix A.

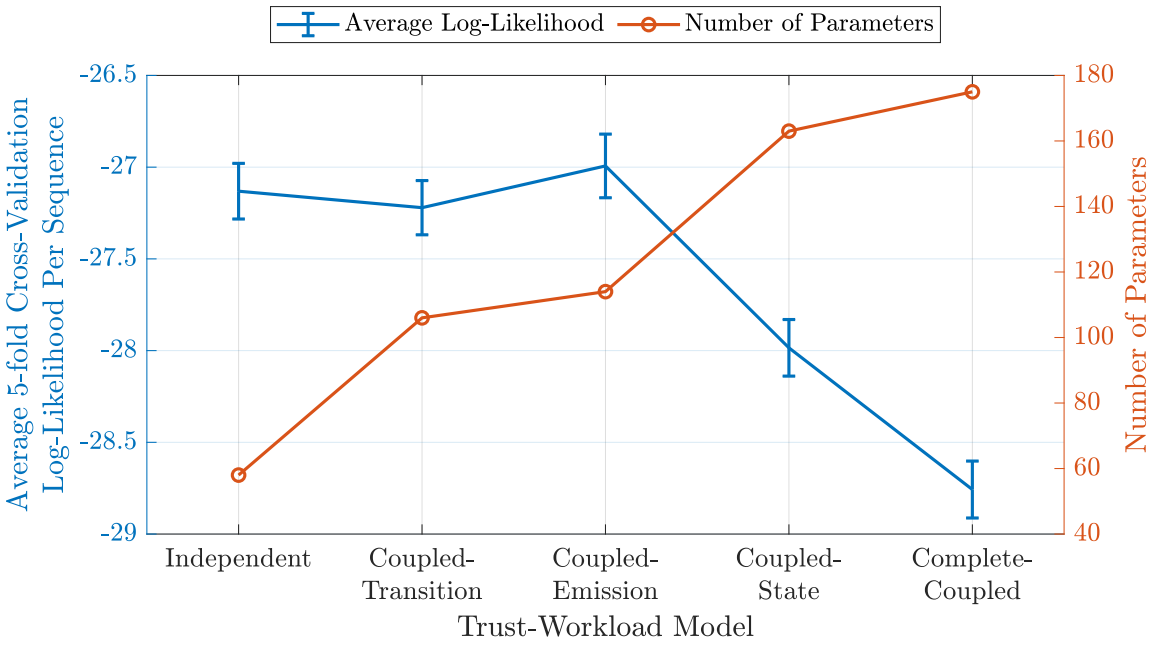


Figure 4.6. Average five-fold cross-validation log-likelihood and number of parameters of the models for ten iterations. Error bars represent the standard error of the mean accuracy across ten iterations and five folds.

Here, we present some key observations about the estimated coupled models. Note that the independent models have been discussed in detail in Chapter 3.

4.3.1 Coupled-Transition Model

The coupled-transition model represented in Figure 4.2 consists of two coupled POMDP models: a trust model and a workload model, which interact in their transition probabilities. When compared to the independent model, the coupled-transition model consists of a similar structure for the observation probability functions for trust and workload.

Considering the trust model, we observe that the emission probability function (Figure 4.7(b)) for the coupled-transition model has relatively similar values to that of the independent model (Figure 4.7(a)). The probability of the human complying with

the automation's recommendation when they are in the state of High Trust is approximately the same for the independent model (0.9787) and the coupled-transition model (0.9805). Interestingly, the probability of the human not complying with the automation's recommendation when they are in the state of Low Trust is slightly different (1.0000 vs. 0.8411); nonetheless, both probabilities are relatively high. Therefore, in both cases, the models suggest that it is highly likely that the human will not comply with the automation's recommendation if they are in the Low Trust state.

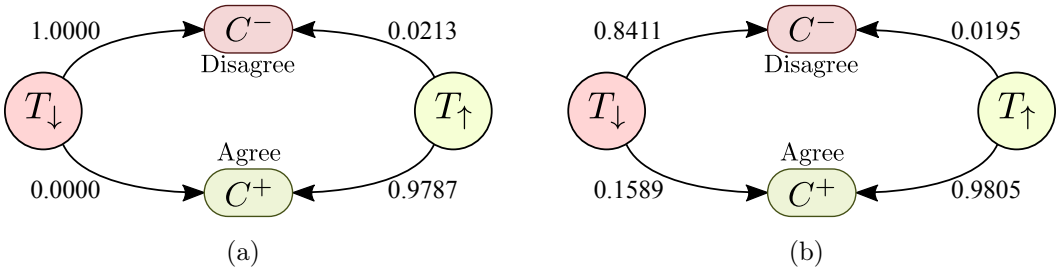


Figure 4.7. Emission probability function $\mathcal{E}_T(o_C|s_T)$ for trust in the independent and coupled-transition model. The left diagram (a) shows the emission probability function for the independent model and the right diagram (b) shows the emission probability function for the coupled-transition model. Probabilities of observation are shown beside the arrows.

We now compare the trust transition probability function of the coupled-transition model in the case when the human was previously in a state of Low Workload $W_\downarrow$ (Figure 4.8) to that when the human was previously in a state of High Workload $W_\uparrow$ (Figure 4.9). Recall that in contrast to the independent model, the trust transition probability function for the coupled-transition model is not only dependent on the previous state of trust and the action, but it also depends on the previous state of workload. In the high risk case (i.e., when the recommendation suggests Light Armor $a_{S_A} = S_A^-$), the probability of transitioning to a state of High Trust $T_\uparrow$ from any state of trust is higher for the High Workload $W_\uparrow$ case than for the Low Workload $W_\downarrow$ case for a given transparency and experience (compare Figure 4.9(a) with Figure 4.8(a)

and Figure 4.9(b) against Figure 4.8(b)). Note that the decision aid suggesting Light Armor represents a high risk situation for over-trust because an incorrect human decision of complying with the recommendation can lead to the human using Light Armor in the presence of gunmen, resulting in injury and an extra penalty of 20 seconds. Nonetheless, for the low risk case (i.e., when the recommendation suggests Heavy Armor $a_{S_A} = S_A^+$), the trust transition probabilities are almost similar between the High Workload $W_\uparrow$ case and the Low Workload $W_\downarrow$ case for a given transparency and experience (compare Figure 4.9(c) with Figure 4.8(c) and Figure 4.9(d) with Figure 4.8(d)). This means that the human's trust is higher for the high workload situation as compared to the low workload situation when the risk is high. This observations is consistent with the recent finding in [193] that trust is comparable between high and low workload conditions, but higher risk elevates trust in high workload conditions. In [193], the authors observe that humans have higher levels of trust when 1) in a multitasking environment that demands greater attention and 2) their perceived risk is high, regardless of the true reliability of automated systems. Findings in [194] also suggest that trust in automation declines when the primary task demands more attention. Therefore, the coupled-transition model is successfully able to capture these nuanced effects of workload on human trust dynamics.

Considering the workload model, we observe that the emission probability function (Figure 4.11) for the coupled-transition model has similar values to that of the independent model (Figure 4.10). Therefore, the probability density functions (PDFs) of human response time, given a state of workload, are essentially the same for the independent model and the coupled-transition model. Again, similar to the trust transition probability function, the workload transition probability function for the coupled-transition model is not only dependent on the previous state of workload and the action, but it also depends on the previous state of trust. As with the trust model, the coupled-transition model also captures subtle differences between the transition probabilities of workload states for low and high trust states. Please refer to Appendix A for a more detailed description of the coupled-transition model.

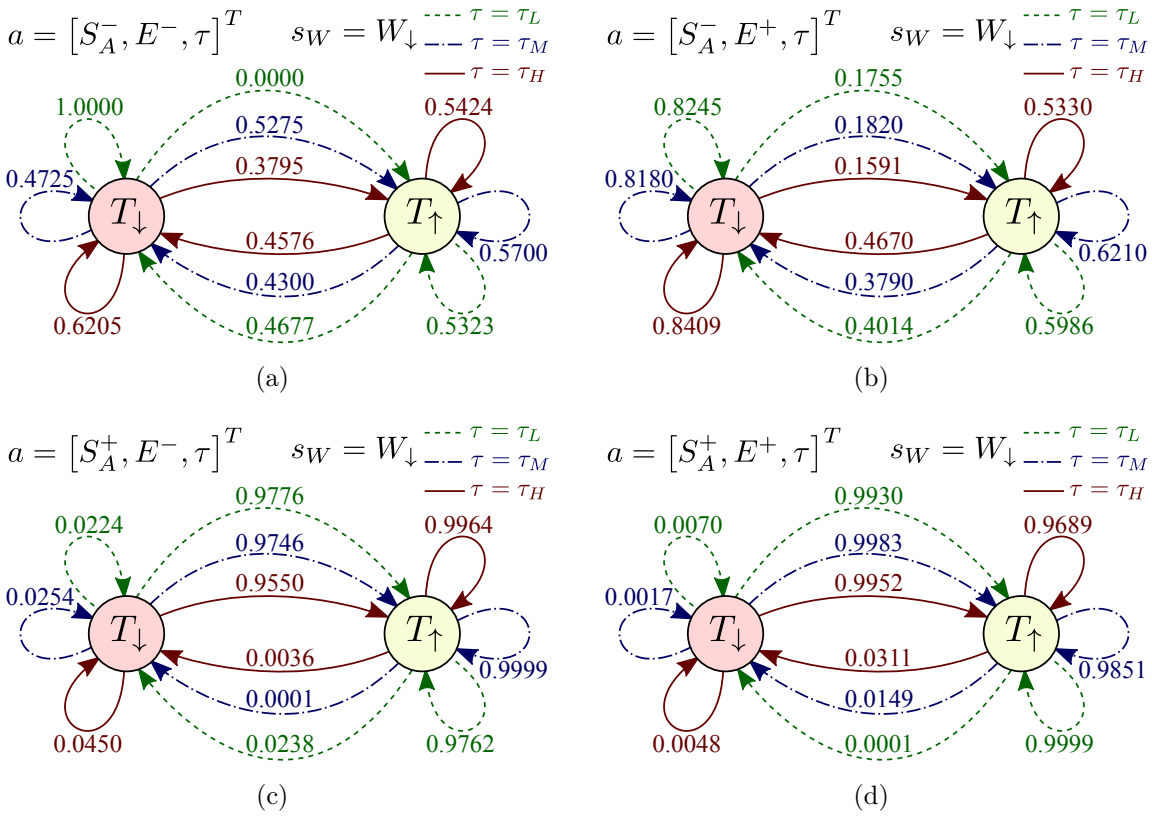


Figure 4.8. Transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\downarrow, a)$ for trust in the coupled-transition model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid's recommendation is Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

4.3.2 Coupled-Emission Model

The coupled-emission model represented in Figure 4.3 considers the interaction between compliance and workload as well as response time and trust in the emis-

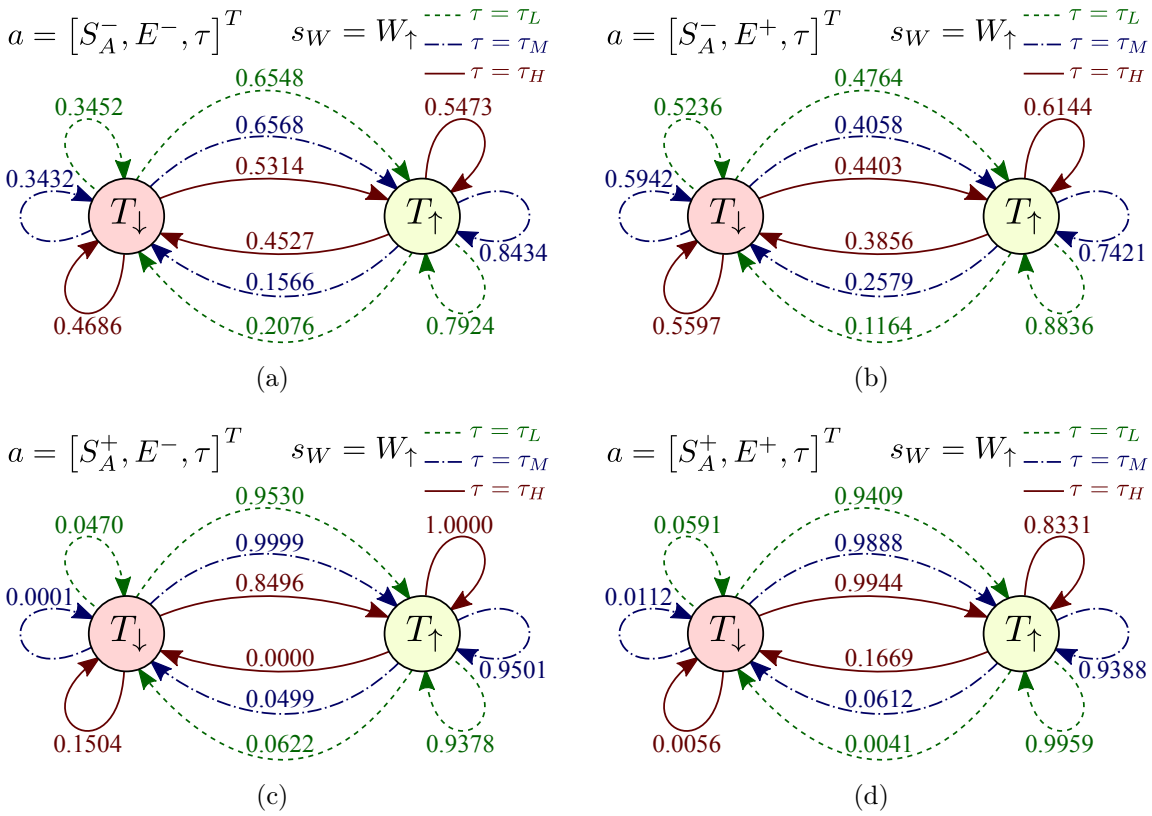


Figure 4.9. Transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\uparrow, a)$ for trust in the coupled-transition model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid's recommendation is Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

sion probability functions, apart from the interactions in the transition probability functions as discussed in the coupled-transition model. Therefore, the emission probability of compliance is dependent on both the trust and workload states, and the emission probability of response time is also dependent on both states.

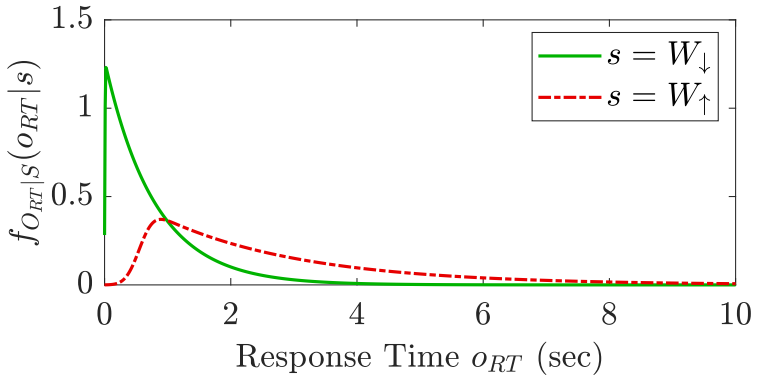


Figure 4.10. Emission probability function $\mathcal{E}_W(o_{RT}|s_W)$ for workload in the independent model. For Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_{\downarrow}}(o_{RT}|W_{\downarrow})$ is characterized by an ex-Gaussian distribution with $\mu_{W_{\downarrow}} = 0.0047$, $\sigma_{W_{\downarrow}} = 0.0062$, and $\tau_{W_{\downarrow}} = 0.7917$. For High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_{\uparrow}}(o_{RT}|W_{\uparrow})$ is characterized by an ex-Gaussian distribution with $\mu_{W_{\uparrow}} = 0.5581$, $\sigma_{W_{\uparrow}} = 0.1745$, and $\tau_{W_{\uparrow}} = 2.2544$.

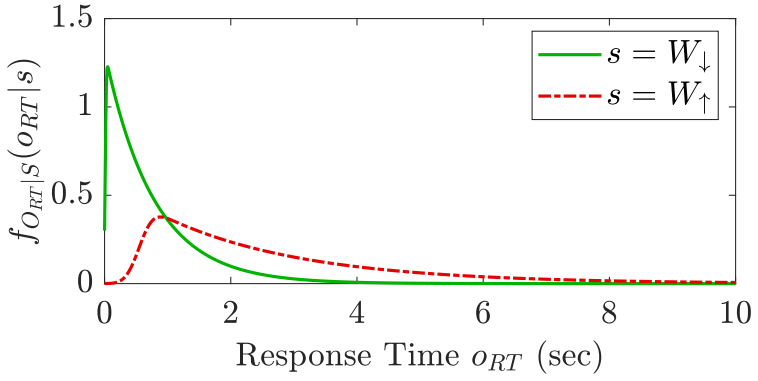


Figure 4.11. Emission probability function $\mathcal{E}_W(o_{RT}|s_W)$ for workload in the coupled-transition model. For Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_{\downarrow}}(o_{RT}|W_{\downarrow})$ is characterized by an ex-Gaussian distribution with $\mu_{W_{\downarrow}} = 0.0108$, $\sigma_{W_{\downarrow}} = 0.0149$, and $\tau_{W_{\downarrow}} = 0.7708$. For High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_{\uparrow}}(o_{RT}|W_{\uparrow})$ is characterized by an ex-Gaussian distribution with $\mu_{W_{\uparrow}} = 0.5566$, $\sigma_{W_{\uparrow}} = 0.1717$, and $\tau_{W_{\uparrow}} = 2.2179$.

We now consider the emission probability functions for the estimated model. Comparing emission probabilities for compliance for Low Workload $W_{\downarrow}$ (Figure 4.12(a)) to that for High Workload $W_{\uparrow}$ (Figure 4.12(b)), we see that participants are more likely to comply with the recommendation in the High Workload state even when they are in a state of Low Trust, as compared when they are in a Low Workload state. This observation agrees with the findings in [66] that suggest compliance is higher in high workload situations. Similarly, comparing the response time emission probability function for the Low Trust state $T_{\downarrow}$ (Figure 4.13(a)) to that for the High Trust state $T_{\uparrow}$ (Figure 4.13(b)), we observe that High Trust has a higher probability of faster response time than Low Trust even for the state of High Workload. It means that the participants responded faster when their trust was high even with higher workload. Further details about this model is presented in Appendix A.

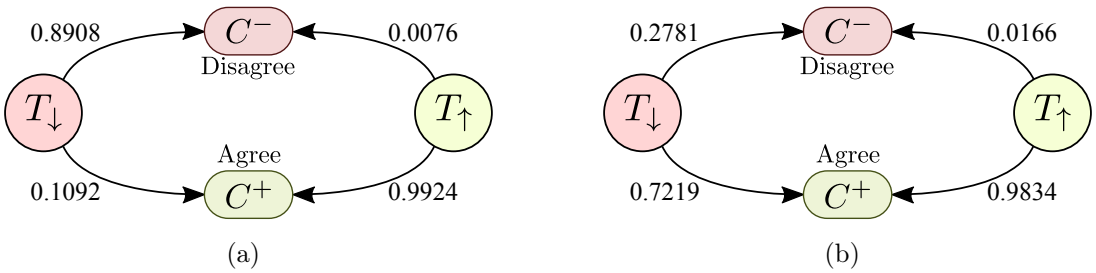


Figure 4.12. Emission probability function $\mathcal{E}_T(o_C|s_T, s_W)$ for trust in the coupled-emission model. Probabilities of observation are shown beside the arrows. The left diagram (a) shows the emission probabilities when the workload state is $W_{\downarrow}$. The right diagram (b) shows the emission probabilities when the workload state is $W_{\uparrow}$.

In summary, the proposed coupled models provide a rich framework for capturing the dynamics of human trust and workload behavior while also considering the subtle interactions between them. These models are able to quantitatively model these interactions, which have been discussed only qualitatively in the existing literature. In the next section, we leverage these estimated models to calculate optimal control

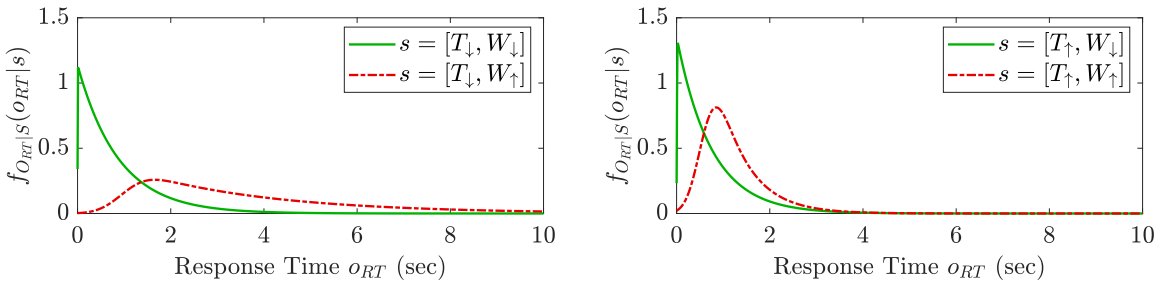


Figure 4.13. Emission probability function $\mathcal{E}_W(o_{RT}|s_T, s_W)$ for workload in the coupled-emission model. The left diagram (a) shows the emission probabilities when the trust state is $T_\downarrow$. For Low Trust and Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|T_\downarrow,W_\downarrow}(o_{RT}|T_\downarrow, W_\downarrow)$ is characterized by an ex-Gaussian distribution with $\mu_{T_\downarrow,W_\downarrow} = 0.0018$, $\sigma_{T_\downarrow,W_\downarrow} = 0.0034$, and $\tau_{T_\downarrow,W_\downarrow} = 0.8804$. For Low Trust and High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|T_\downarrow,W_\uparrow}(o_{RT}|T_\downarrow, W_\uparrow)$ is characterized by an ex-Gaussian distribution with $\mu_{T_\downarrow,W_\uparrow} = 0.9845$, $\sigma_{T_\downarrow,W_\uparrow} = 0.4138$, and $\tau_{T_\downarrow,W_\uparrow} = 2.8825$. The right diagram (b) shows the emission probabilities when the trust state is $T_\uparrow$. For High Trust and Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|T_\uparrow,W_\downarrow}(o_{RT}|T_\uparrow, W_\downarrow)$ is characterized by an ex-Gaussian distribution with $\mu_{T_\uparrow,W_\downarrow} = 0.0063$, $\sigma_{T_\uparrow,W_\downarrow} = 0.0067$, and $\tau_{T_\uparrow,W_\downarrow} = 0.7439$. For High Trust and High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|T_\uparrow,W_\uparrow}(o_{RT}|T_\uparrow, W_\uparrow)$ is characterized by an ex-Gaussian distribution with $\mu_{T_\uparrow,W_\uparrow} = 0.5578$, $\sigma_{T_\uparrow,W_\uparrow} = 0.2603$, and $\tau_{T_\uparrow,W_\uparrow} = 0.6510$.

policies and conduct closed-loop validation using real-time trust and workload state estimation.

4.4 Model Validation and Results

To experimentally validate and compare the performance of the estimated independent and coupled models, we calculate the control policies for each of the three models using the reward functions presented in Section 3.4 with $\zeta = 0.50$, $\zeta = 0.85$, and $\zeta = 0.95$. Note that the reward functions with higher values of ζ have higher weights for decision rewards (which penalize incorrect decisions) than response time

rewards (which penalize slower response times). The control policies are evaluated using the methodology described in Section 3.4.3. The control policies are presented in Appendix B. Using these control policies, we conducted four human subject studies. These experiments are identical to the one used to collect open-loop data for each transparency but with transparency now controlled using the control policies based on human trust and workload estimation.

4.4.1 Stimuli and Procedure:

Four within-subject studies were performed in which participants were asked to interact with a simulation of three reconnaissance missions as described in Section 3.2.2. However, instead of fixed transparency in each mission, the transparency was controlled based on a feedback control policy for some missions. The details of the four studies are summarized in Table 4.1. For the first study, low transparency was always used in one of the three missions, and in the other two missions, the transparency was dynamically varied based on control policies corresponding to $\zeta = 0.50$ for the independent model and coupled-transition model, respectively. In the second study, medium transparency was always used in one of the three missions, and in the other two missions, the transparency was dynamically varied based on control policies corresponding to $\zeta = 0.85$ for the independent model and coupled-transition model, respectively. In the third study, high transparency was always used in one of the three missions, and in the other two missions, the transparency was dynamically varied based on control policies corresponding to $\zeta = 0.95$ for the independent model and coupled-transition model, respectively. Finally, in the fourth study, in the three missions, the transparency was dynamically varied based on the control policy corresponding to $\zeta = 0.50$, $\zeta = 0.85$, and $\zeta = 0.95$, respectively, for the coupled-emission model. Three missions in each study ensured that the studies were short enough to avoid participant fatigue and were consistent in structure with

the study used to collect open-loop data. Moreover, the order of missions was again randomized across participants to reduce ordering effects [185].

Table 4.1.
Summary of the four closed-loop studies used to compare the performance of the independent and the coupled models of interest.

	Control policy in the three missions for varying transparency	Number of participants
Study 1	• Fixed low transparency • Independent model with $\zeta = 0.50$ • Coupled-transition model with $\zeta = 0.50$	• Total: 56 • Outlying: 9 • Remaining: 47
Study 2	• Fixed medium transparency • Independent model with $\zeta = 0.85$ • Coupled-transition model with $\zeta = 0.85$	• Total: 52 • Outlying: 3 • Remaining: 49
Study 3	• Fixed high transparency • Independent model with $\zeta = 0.95$ • Coupled-transition model with $\zeta = 0.95$	• Total: 54 • Outlying: 10 • Remaining: 44
Study 4	• Coupled-emission model with $\zeta = 0.50$ • Coupled-emission model with $\zeta = 0.85$ • Coupled-emission model with $\zeta = 0.95$	• Total: 53 • Outlying: 5 • Remaining: 48

4.4.2 Participants

Fifty-six participants for the first study, fifty-two participants for the second study, fifty-four participants for the third study, and fifty-three participants for the fourth study, participated in and completed the study online. They were recruited using Amazon Mechanical Turk [97] with the same criteria used for the earlier study. To

account for participants who were not sufficiently engaged in the study, we filtered out data that had any response time higher than 40.45 seconds (the 99.5 percentile value of all response times for the open-loop study data). The number of participants removed from the datasets as a result of this filtering is shown in Table 4.1.

4.4.3 Decision Reward and Response Time Reward

Using the data collected from the four validation studies, we quantify and evaluate the participants' performance for the fixed transparency missions and the dynamically varying transparency missions for different models. We first compare two metrics: total decision reward and total response time reward for each type of mission. We use linear mixed effects analysis and likelihood ratio tests to determine whether the use of trust-workload behavior-based feedback has any significant effect on these metrics. We use the statistical computing language R [189] and *lme4* library [190] to perform a linear mixed effect model approach to analyze the relationship between each of the metrics and transparency policies. As a fixed effect, we use the transparency policy in the models. To account for variations in the metrics calculated for different participants, the models considered each individual as a random effect. P-values are obtained using likelihood ratio tests of the full model that includes the transparency policy as a fixed effect against the model that does not include the transparency policy. Figure 4.14 shows the effect of the transparency policies (open-loop: Low, Medium, and High; closed-loop: independent model, coupled-transition model, and coupled-emission model with $\zeta = 0.50, 0.85,$ and 0.95, respectively) on the total decision reward and on the total response time reward across participants.

The total decision reward is defined as the sum of all decision rewards, based on Table 3.5, accrued by the participant in a mission. A likelihood ratio test using linear mixed effects models indicates that the transparency policies significantly affected total decision reward ($\chi^2(11) = 40.193, p \approx 0.0000$). The total response time reward is defined as the negative of the sum of all response times, in seconds, accrued by

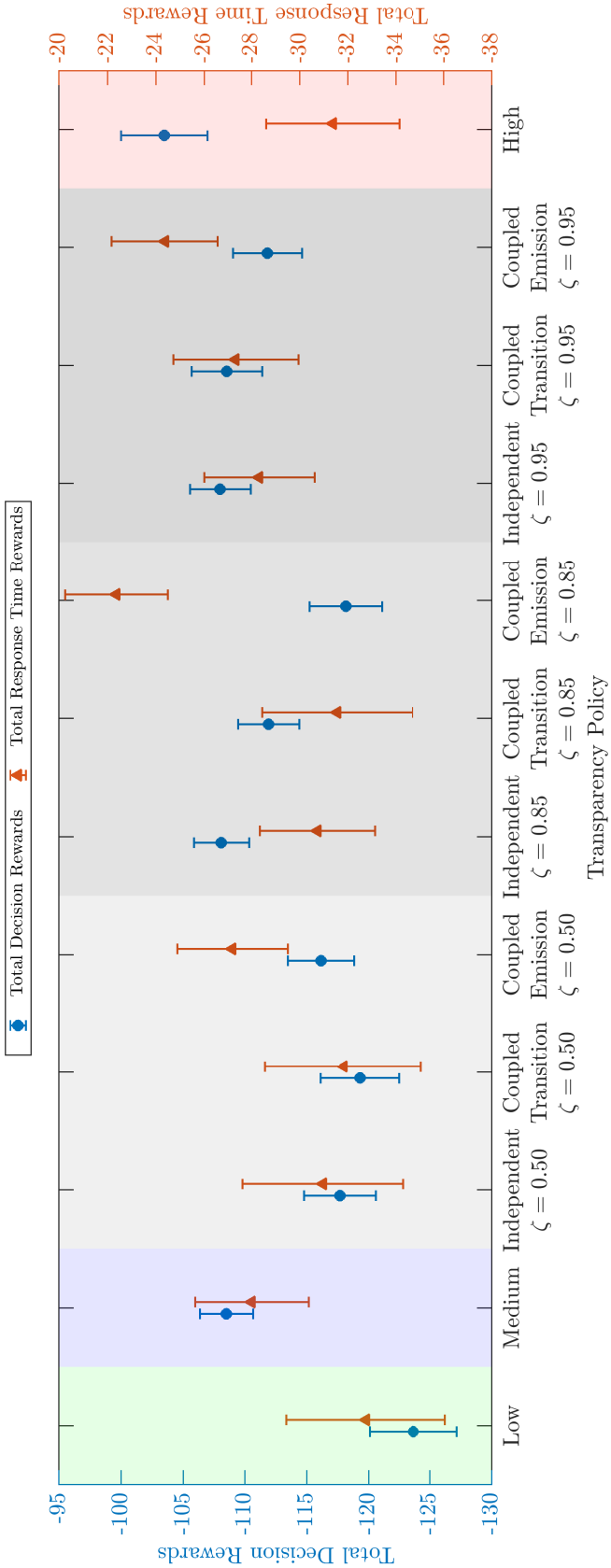


Figure 4.14. Effect of the control policies on the total decision and total response time rewards. Error bars represent the standard error of the mean across participants. The closed-loop control policies are highlighted in gray.

the participant in a mission. A likelihood ratio test indicates that the transparency policies did not significantly affect total response time reward ($\chi^2(11) = 14.932, p = 0.1856$). Nonetheless, there are some trends in the mean response time reward across the control policies.

Considering open-loop fixed transparency policies, we see from Figure 4.14 that high transparency has the best performance in terms of decision rewards, followed by medium and low transparency. Furthermore, control policies with higher values of ζ achieve a higher decision reward for a given model. This is expected because with higher values of ζ, the control policies have a higher weight to maximize decision rewards. Interestingly, the independent model based policies achieve a higher (if not equal) decision reward than that of coupled models for a given value of ζ.

In terms of response time rewards, considering open-loop fixed transparency policies, we observe that medium transparency has the highest mean value. This is possibly because high transparency requires the participant to process more information, which takes more time, and conversely, low transparency does not provide enough information to the participants, thereby leading to confusion and a longer response time. For the closed-loop policies, the response time rewards are not significantly different across different values of ζ. However, we observe that the control policies based on the coupled models, in particular, the coupled-emission model, achieves a better (if not equal) response time reward for a given value of ζ. This is possibly because the coupled models capture the effect of both trust and workload on response time.

4.4.4 Total Reward

To objectively compare the performance of the independent and coupled models, we compare the total reward metric between the models from the collected closed-loop

data for each value of ζ. The total reward is the actual value of the reward function attained in the closed-loop study (see Section 3.4) and is defined as

$$\mathcal{R} = \zeta \mathcal{R}_T + (1 - \zeta)\mathcal{R}_W \quad .$$

Note that the control policies were calculated such that they maximize the reward function. Therefore, the model whose control policy is able to attain a higher value of total reward has an objectively better performance for a given ζ.

Again, we use linear mixed effects analysis and likelihood ratio tests as discussed earlier to determine whether the control policies based on the different models have any significant effect on the total reward for each ζ. Figure 4.15 shows the effect of the control policies, based on the independent and coupled models, on the total reward across participants for $\zeta = 0.50$. A likelihood ratio test using linear mixed effects models indicates that the control policies do not have a statistically significant effect on the total reward for $\zeta = 0.50$ ($\chi^2(2) = 1.5029, p = 0.4717$). Nonetheless,

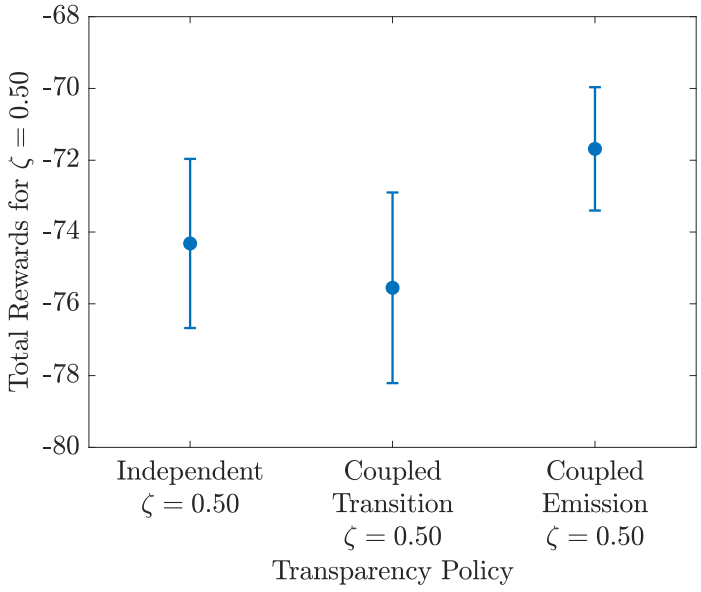


Figure 4.15. Effect of the control policies based on the independent and coupled models on the total reward for $\zeta = 0.50$. Error bars represent the standard error of the mean across participants.

based on the trends, the coupled-emission model achieves slightly better performance than the independent model. Figure 4.16 shows the effect of the control policies based on the independent and coupled models on the total reward across participants for $\zeta = 0.85$. A likelihood ratio test using linear mixed effects models indicates that the

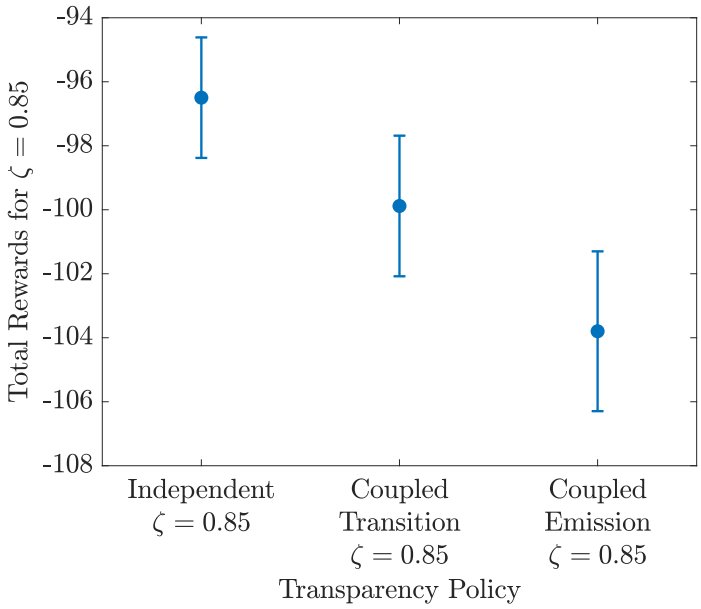


Figure 4.16. Effect of the control policies on the total reward based on the independent and coupled models for $\zeta = 0.85$. Error bars represent the standard error of the mean across participants.

control policies have a statistically significant effect on the total reward for $\zeta = 0.85$ ($\chi^2(2) = 5.4904, p = 0.06423$). We see that the independent model outperforms the coupled models.

Figure 4.17 shows the effect of the control policies on the total reward across participants based on the independent and coupled models for $\zeta = 0.95$. A likelihood ratio test using linear mixed effects models indicates that the control policies do not have a statistically significant effect on the total reward for $\zeta = 0.95$ ($\chi^2(2) = 1.0253, p = 0.5989$). Nonetheless, we see that the coupled models result in a lower, if not equal, total reward than the independent model.

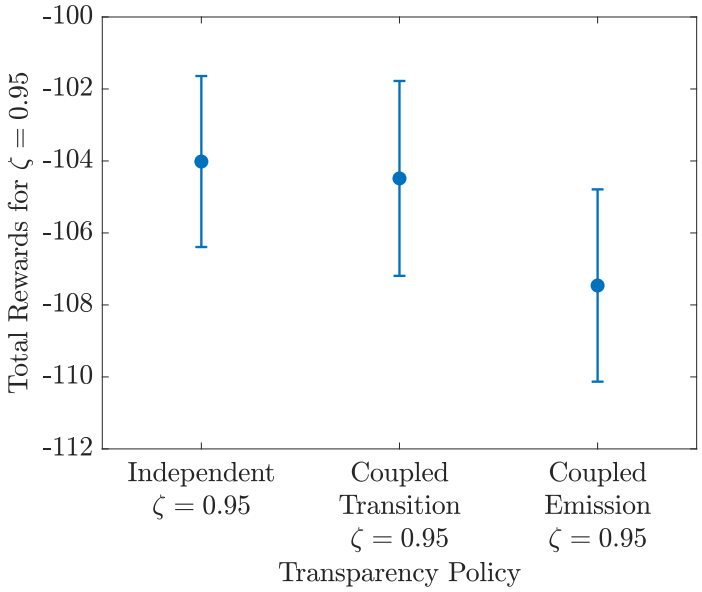


Figure 4.17. Effect of the control policies based on the independent and coupled models on the total reward for $\zeta = 0.95$. Error bars represent the standard error of the mean across participants.

Based on the total rewards, we conclude that there is possibly no explicit benefit in terms of improved closed-loop performance with the use of coupled models. In particular, for improved decision making, which is obtained using a higher value of ζ, the independent model performs better, if not equal, than coupled models. This is possibly because the human's decisions are directly related to human trust behavior and are relatively less influenced by workload. However, in time-critical contexts where decision time is just as important as making the correct decision, closed-loop policies based on the coupled-emission model can achieve better performance as they potentially capture the interaction dynamics between human trust and workload.

4.5 Chapter Summary

In this chapter, we modeled the coupling between human trust and workload in a context involving a human interacting with an automated decision-aid. We explored

and analyzed multiple models with varying complexities by relaxing assumptions on trust and workload independence. We found that the proposed coupled models were successfully able to capture the nuanced interactions between human trust and workload dynamics. Finally, we compared the performance of two of the coupled models to that of the independent model by validating the optimal control policy for dynamically varying automation transparency based on each model's trust and workload estimates. We concluded that there is possibly no explicit benefit in terms of improved closed-loop performance with the use of coupled models, except in time-critical contexts. It should be noted that with the increase in complexity of the coupled models, the sample size of the data required for parameter estimation also increases. Additionally, the increased complexity of the model reduces the model interpretability based on the parameter values. Therefore, one should consider this tradeoff while employing coupled models of trust and workload.

The model evaluation presented here could be improved by addressing a few limitations. The models were trained using data from 196 participants; a larger sample size may potentially improve the performance of the models. Additionally, while we assume that the trust-workload behavior of the population can be captured using one model, clustering algorithms could be used to determine whether there are fundamental behavioral differences in the population. Nevertheless, the framework presented here provides a substantial step forward toward the development of quantitative dynamic models of human behavior and their use for implementing adaptive automation in human-machine interaction contexts.

5. CONCLUSIONS

5.1 Summary of Research Contributions

Interactions between humans and automation can be improved by designing automation that can infer human behavior and respond accordingly. In this dissertation, I developed a framework to model the dynamics of human trust and workload as they evolve during a human's interaction with a decision-aid system. I further designed and validated a model-based feedback control policy aimed at dynamically varying the automation's transparency to improve the overall performance of the human-machine team. More specifically, I developed both a classical state-space model and machine learning model to quantitatively predict and estimate human trust in real time. I showed how some of these models can be combined to predict human trust based on a combination of human behavioral and psychophysiological measurements. Thereafter, I developed a probabilistic dynamic model to capture the dynamics of human trust along with human workload. I used this model to synthesize optimal control policies and validated the policies in closed-loop using human subject experiments. Finally, I explored and analyzed the dynamic coupling between human trust and workload to strengthen the model framework.

While developing my state-space model of trust, I identified the significance of cumulative trust and expectation bias through experiments that elicited multiple dynamic transitions in human trust, and then incorporated these two variables in the proposed linear model. In addition to proposing a general trust model structure, I characterized the effects of both dispositional and learned trust factors, specifically national culture, gender and system error type, using estimated model parameters. I also characterized the effects of misses and false alarms on the dynamics of human trust behavior and compared differences between demographics. While the proposed

model is representative of a population of individuals rather than trained to a specific human, such a model can be used to design machines that are required to interact with unspecified users grouped by demographics.

For the machine learning model of human trust, I developed two approaches for classifier-based empirical trust sensor models that estimate human trust level using psychophysiological measurements. The first approach was to consider a common set of psychophysiological features as the input variables for any human and train a classifier-based model using this feature set, resulting in a general trust sensor model with a mean accuracy of 71.22%. The second approach was to consider a customized feature set for each individual and train a classifier-based model using that feature set; this resulted in a mean accuracy of 78.55%. The primary trade-off between these two approaches is between training time and performance (based on mean accuracy) of the classifier-based model. Later, I demonstrated an approach to incorporate behavioral dynamics in these static classification algorithms.

Next I presented a partially observable Markov decision process (POMDP) model for human trust and workload. The model, which was parameterized using human subject data, captured the effects of the decision-aid's recommendation, the human's previous experience with the automation, and automation transparency on the human's trust-workload behavior. The model is capable of estimating human trust and workload in real time using recursive belief-state estimates. Furthermore, experimental validation showed that the closed-loop control policies were successfully able to manage the human decision versus response time performance tradeoff based on a tuning parameter in the reward function. At last, I also extended the framework to explore and analyze the coupling interactions between human trust and workload. I found that although there is no explicit closed-loop performance benefit of modeling the coupling between the two states, the coupled model does captures some nuanced characteristics of trust and workload interactions. My proposed framework provides a tractable methodology for using human behavior as a real-time feedback signal to optimize human-machine interactions through dynamic modeling and control.

5.2 Future Research Directions

This dissertation presented a human state-based feedback framework that focused on human trust and workload during an interaction with an automation in a decision-aid context. However the model framework also lays the groundwork for multiple impactful future research directions. Some of these directions should include (1) validating the performance of the framework while combining both psychophysiological and behavioral measurements to estimate human states, (2) extending and validating the proposed framework in other contexts, (3) customizing the framework to capture individual differences between humans, and (4) augmenting the framework to include other control variables to influence human states such as control authority. Each of these directions would not only expand the applicability of the presented framework, but also demonstrate its efficacy in designing human-aware automation.

With the developed framework for the human trust sensor model using real-time psychophysiological measurements, we can estimate human trust level using psychophysiological data even in the absence of the behavioral data required for POMDP model-based trust estimation. However, the estimated trust from psychophysiological measurements still needs to be incorporated in the closed-loop framework. In the absence of either of the two trust estimates (psychophysiological measurements-based or behavioral data-based), the available estimate of human trust can be used in the closed-loop framework described in Chapter 3.4. Moreover, in situations when both trust estimates are available, a combined trust estimate could be more robust to uncertainties in the environment and in the nature of the human-machine interaction itself. In this direction, further validation is required to evaluate the performance and robustness of the framework.

Moreover, while this dissertation focused on a case study considering a reconnaissance mission task, the model framework could be used in a variety of other decision-aid contexts, such as health recommender systems and other assistive robot applications, by retraining the model using new context-specific data. Future work could con-

sider extensions of the framework to action automation by re-defining context-specific observations, actions, and reward function(s). Furthermore, since a computer-based simulated interface was used in the experiment, the ecological validity could be improved by testing the established framework in both immersive environments, for example, flight or driving simulators, as well as real-life settings.

The presented framework further assumes a single model for the general population, even though the model and the corresponding optimal control policy might also depend on factors based on the demographics of each individual. Future work could carefully investigate the effect of these individual-specific factors. Furthermore, another research direction could be identifying clusters of people with similar trust-workload behavior and creating customized models for each cluster.

Finally, we only used automation transparency to influence human behavior. However, other variables such as the amount of control the automation shares with human and degree of conservativeness of the automation in a risky situation also impact human behavior. Therefore, effects of these variables along with other context-specific variables can also be modeled; thereafter, these variables can be optimally controlled to improve human-machine interactions.

REFERENCES

[1] Y. Wang and F. Zhang, Eds., *Trends in Control and Decision-Making for Human–Robot Collaboration Systems*. Cham: Springer International Publishing, 2017.

[2] N. R. Jennings, L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, and A. Rogers, "Human-agent Collectives," *Communications of the ACM*, vol. 57, no. 12, pp. 80–88, Nov. 2014.

[3] B. M. Muir, "Trust between humans and machines, and the design of decision aids," *International Journal of Man-Machine Studies*, vol. 27, no. 5–6, pp. 527–539, 1987.

[4] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 46, no. 1, pp. 50–80, 2004.

[5] T. B. Sheridan and R. Parasuraman, "Human-automation interaction," *Reviews of human factors and ergonomics*, vol. 1, no. 1, pp. 89–129, 2005.

[6] M. Richtel and C. Dougherty, "Google's driverless cars run into problem: Cars with drivers," *New York Times*, vol. 1, 2015.

[7] J. Y. Chen, K. Procci, M. Boyce, J. Wright, A. Garcia, and M. Barnes, "Situation awareness-based agent transparency," Army Research Lab Aberdeen Proving Ground MD Human Research and Engineering Directorate, Tech. Rep., 2014.

[8] J. E. Mercado, M. A. Rupp, J. Y. C. Chen, M. J. Barnes, D. Barber, and K. Procci, "Intelligent Agent Transparency in Human–Agent Teaming for Multi-UxV Management," *Hum Factors*, vol. 58, no. 3, pp. 401–415, May 2016.

[9] T. Helldin, "Transparency for future semi-automated systems: Effects of transparency on operator performance, workload and trust," Ph.D. dissertation, Örebro University, Örebro, 2014.

[10] N. Lyu, L. Xie, C. Wu, Q. Fu, and C. Deng, "Driver's cognitive workload and driving performance under traffic sign information exposure in complex environments: A case study of the highways in China," *International journal of environmental research and public health*, vol. 14, no. 2, p. 203, 2017.

[11] J.-H. Cho, K. Chan, and S. Adali, "A Survey on Trust Modeling," *ACM Comput. Surv.*, vol. 48, no. 2, pp. 28:1–28:40, Oct. 2015.

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

[13] R. Parasuraman, T. Sheridan, and C. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, May 2000.

[14] R. Parasuraman and C. D. Wickens, "Humans: Still Vital After All These Years of Automation," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 50, no. 3, pp. 511–520, Jun. 2008.

[15] C. D. Wickens and S. R. Dixon, "The benefits of imperfect diagnostic automation: A synthesis of the literature," *Theoretical Issues in Ergonomics Science*, vol. 8, no. 3, pp. 201–212, May 2007.

[16] P. A. Hancock, M. H. Chignell, and A. Lowenthal, "An adaptive human-machine system," in *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, vol. 15, 1985, pp. 627–629.

[17] W. B. Rouse, "Adaptive Aiding for Human/Computer Control," *Hum Factors*, vol. 30, no. 4, pp. 431–443, Aug. 1988.

[18] R. Parasuraman, T. Bahri, J. E. Deaton, J. G. Morrison, and M. Barnes, "Theory and Design of Adaptive Automation in Aviation Systems," Catholic Univ of America Washington DC Cognitive Science Lab, Tech. Rep., Jul. 1992.

[19] V. Alonso and P. de la Puente, "System Transparency in Shared Autonomy: A Mini Review," *Front. Neurorobot.*, vol. 12, p. 83, Nov. 2018.

[20] N. Tintarev and J. Masthoff, "A Survey of Explanations in Recommender Systems," in *2007 IEEE 23rd International Conference on Data Engineering Workshop*, Apr. 2007, pp. 801–810.

[21] A. Felfernig and B. Gula, "An Empirical Study on Consumer Behavior in the Interaction with Knowledge-based Recommender Applications," in *The 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services (CEC/EEE'06)*, Jun. 2006, pp. 37–37.

[22] R. Sinha and K. Swearingen, "The Role of Transparency in Recommender Systems," in *CHI'02 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2002, p. 2.

[23] N. Wang, D. V. Pynadath, and S. G. Hill, "The impact of POMDP-generated explanations on trust and performance in human-robot teams," in *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2016, pp. 997–1005.

[24] ——, "Trust calibration within a human-robot team: Comparing automatically generated explanations," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*. IEEE, 2016, pp. 109–116.

[25] N. Wang, D. V. Pynadath, K. V. Unnikrishnan, S. Shankar, and C. Merchant, "Intelligent Agents for Virtual Simulation of Human-Robot Interaction," in *Virtual, Augmented and Mixed Reality*, R. Shumaker and S. Lackey, Eds. Cham: Springer International Publishing, 2015, vol. 9179, pp. 228–239.

[26] M. Hosseini, A. Shahri, K. Phalp, and R. Ali, "Four reference models for transparency requirements in information systems," *Requirements Engineering*, vol. 23, no. 2, pp. 251–275, Jun. 2018.

[27] M. Ananny and K. Crawford, "Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability," *New Media & Society*, vol. 20, no. 3, pp. 973–989, Mar. 2018.

[28] L. Bohua, S. Lishan, and R. Jian, "Driver's visual cognition behaviors of traffic signs based on eye movement parameters," *Journal of Transportation Systems Engineering and Information Technology*, vol. 11, no. 4, pp. 22–27, 2011.

[29] H. Makishita and K. Matsunaga, "Differences of drivers' reaction times according to age and mental workload," *Accident Analysis & Prevention*, vol. 40, no. 2, pp. 567–575, Mar. 2008.

[30] R. P. Heitz, "The speed-accuracy tradeoff: History, physiology, methodology, and behavior," *Front. Neurosci.*, vol. 8, 2014.

[31] J. I. Gold and M. N. Shadlen, "The Neural Basis of Decision Making," *Annu. Rev. Neurosci.*, vol. 30, no. 1, pp. 535–574, Jun. 2007.

[32] J. Drugowitsch, G. C. DeAngelis, D. E. Angelaki, and A. Pouget, "Tuning the speed-accuracy trade-off to maximize reward rate in multisensory decision-making," *eLife*, vol. 4, p. e06678, Jun. 2015.

[33] C. M. Jonker and J. Treur, "Formal Analysis of Models for the Dynamics of Trust Based on Experiences," in *European Workshop on Modelling Autonomous Agents in a Multi-Agent World*. Springer Berlin Heidelberg, 1999, pp. 221–231.

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

[35] R. Croson and N. Buchan, "Gender and culture: International experimental evidence from trust games," *American Economic Review*, vol. 89, no. 2, pp. 386–391, 1999.

[36] T. Nomura and S. Takagi, "Exploring effects of educational backgrounds and gender in human-robot interaction," in *2011 International Conference on User Science and Engineering*, 2011, pp. 24–29.

[37] M. Hoogendoorn, S. W. Jaffry, P. P. Van Maanen, and J. Treur, "Modelling biased human trust dynamics," *Web Intelligence and Agent Systems*, vol. 11, no. 1, pp. 21–40, Aug. 2013.

[38] G. Hofstede, *Culture's Consequences: Comparing Values, Behaviors, Institutions and Organizations Across Nations*, 2nd ed. SAGE Publications, 2001.

[39] R. Riedl and A. Javor, "The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging." *Journal of Neuroscience, Psychology, and Economics*, vol. 5, no. 2, p. 63, 2012.

[40] C. Boudreau, M. D. McCubbins, and S. Coulson, "Knowing when to trust others: An ERP study of decision making after receiving information from unknown people," *Social Cognitive and Affective Neuroscience*, vol. 4, no. 1, pp. 23–34, Nov. 2008.

[41] Y. Long, X. Jiang, and X. Zhou, "To believe or not to believe: Trust choice modulates brain responses in outcome evaluation," *Neuroscience*, vol. 200, pp. 50–58, 2012.

[42] T. C. Handy, *Event-Related Potentials: A Methods Handbook*, ser. A Bradford Book. MIT Press, 2005.

[43] A. Khawaji, J. Zhou, F. Chen, and N. Marcus, "Using Galvanic Skin Response (GSR) to Measure Trust and Cognitive Load in the Text–Chat Environment," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. ACM Press, 2015, pp. 1989–1994.

[44] M. E. G. Moe, M. Tavakolifard, and S. J. Knapskog, "Learning trust in dynamic multiagent environments using HMMs," in *Proceedings of the 13th Nordic Workshop on Secure IT Systems (NordSec 2008)*, 2008.

[45] Z. Malik, I. Akbar, and A. Bouguettaya, "Web services reputation assessment using a hidden Markov model," in *Service-Oriented Computing*. Springer, Berlin, Heidelberg, 2009, pp. 576–591.

[46] K. Akash, W.-L. Hu, T. Reid, and N. Jain, "Dynamic modeling of trust in human-machine interactions," in *American Control Conference (ACC), 2017*. IEEE, 2017, pp. 1542–1548.

[47] W. Hu, K. Akash, T. Reid, and N. Jain, "Computational Modeling of the Dynamics of Human Trust During Human–Machine Interactions," *IEEE Transactions on Human-Machine Systems*, pp. 1–13, 2018.

[48] C. D. Wickens, "Multiple resources and mental workload," *Human factors*, vol. 50, no. 3, pp. 449–455, 2008.

[49] R. Parasuraman, "Designing automation for human use: Empirical studies and quantitative models," *Ergonomics*, vol. 43, no. 7, pp. 931–951, 2000.

[50] M. Hoogendoorn, S. W. Jaffry, P. P. Van Maanen, and J. Treur, "Design and validation of a relative trust model," *Knowledge-Based Systems*, vol. 57, pp. 81–94, Feb. 2014.

[51] J. Sabater and C. Sierra, "Review on computational trust and reputation models," *Artificial intelligence review*, vol. 24, no. 1, pp. 33–60, 2005.

[52] C. Carver and M. Scheier, *Attention and Self-Regulation: A Control-Theory Approach to Human Behavior*, ser. Springer Series in Social Psychology. Springer New York, 2012.

[53] G. Klein, "Naturalistic decision making," *Human factors*, vol. 50, no. 3, pp. 456–460, Jun. 2008.

[54] M. Hoogendoorn, S. W. Jaffry, P.-P. Van Maanen, and J. Treur, "Modeling and validation of biased human trust," in *Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Volume 02*. IEEE Computer Society, 2011, pp. 256–263.

[55] J. A. Colquitt, B. A. Scott, and J. A. LePine, "Trust, trustworthiness, and trust propensity: A meta-analytic test of their unique relationships with risk taking and job performance." *Journal of applied psychology*, vol. 92, no. 4, p. 909, 2007.

[56] G. Ho, L. M. Kiff, T. Plocher, and K. Z. Haigh, "A model of trust & reliance of automation technology for older users," in *AAAI-2005 Fall Symposium:"Caring Machines: AI in Eldercare*, 2005, pp. 45–50.

[57] M. Naef, E. Fehr, U. Fischbacher, J. Schupp, and G. Wagner, "Decomposing trust: Explaining national and ethnical trust differences," Working Paper, Institute for Empirical Research in Economics, University of Zurich, Tech. Rep., 2008.

[58] C. M. Jonker, J. J. P. Schalken, J. Theeuwes, and J. Treur, "Human Experiments in Trust Dynamics," in *Trust Management: Second International Conference, iTrust 2004*. Springer Berlin Heidelberg, 2004, pp. 206–220.

[59] J. D. Lee and N. Moray, "Trust, self-confidence, and operators' adaptation to automation," *International Journal of Human-Computer Studies*, vol. 40, no. 1, pp. 153–184, 1994.

[60] S. Lewandowsky, M. Mundy, and G. P. A. Tan, "The dynamics of trust: Comparing humans to automation." *Journal of Experimental Psychology: Applied*, vol. 6, no. 2, pp. 104–123, 2000.

[61] B. Sadrfaridpour, H. Saeidi, J. Burke, K. Madathil, and Y. Wang, "Modeling and Control of Trust in Human-Robot Collaborative Manufacturing," in *Robust Intelligence and Trust in Autonomous Systems*, R. Mittu, D. Sofge, A. Wagner, and W. Lawless, Eds. Boston, MA: Springer US, 2016, pp. 115–141.

[62] X. Li, T. J. Hess, and J. S. Valacich, "Using attitude and social influence to develop an extended trust model for information systems," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 37, no. 2-3, pp. 108–124, Sep. 2006.

[63] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 697–718, 2003.

[64] H. Poor, *An Introduction to Signal Detection and Estimation*, ser. Springer Texts in Electrical Engineering. Springer New York, 1998.

[65] E. T. Chancey, J. P. Bliss, Y. Yamani, and H. A. H. Handley, "Trust and the Compliance–Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence," *Hum Factors*, vol. 59, no. 3, pp. 333–345, May 2017.

[66] S. R. Dixon and C. D. Wickens, "Automation Reliability in Unmanned Aerial Vehicle Control: A Reliance-Compliance Model of Automation Dependence in High Workload," *Hum Factors*, vol. 48, no. 3, pp. 474–486, Sep. 2006.

[67] J. Meyer, "Effects of warning validity and proximity on responses to warnings." *Human factors*, vol. 43, no. 4, pp. 563–572, Dec. 2001.

[68] S. Breznitz, *Cry Wolf: The Psychology of False Alarms*, 1st ed. Psychology Press, May 2013.

[69] R. Parasuraman and V. Riley, "Humans and automation: Use, misuse, disuse, abuse," *Human factors*, vol. 39, no. 2, pp. 230–253, Jun. 1997.

[70] S. R. Dixon, C. D. Wickens, and J. S. McCarley, "On the independence of compliance and reliance: Are automation false alarms worse than misses?" *Human factors*, vol. 49, no. 4, pp. 564–72, Aug. 2007.

[71] J. P. Bliss and G. Capobianco, "Collective mistrust of alarms," *Proceedings of the International Symposium on Aviation Psychology, April*, pp. 14–18, 2003.

[72] J. D. Johnson, "Type of automation failure: The effects on trust and reliance in automation," Master Thesis, Georgia Institute of Technology, 2004.

[73] C. D. Wickens, S. Rice, D. Keller, S. Hutchins, J. Hughes, and K. Clayton, "False alerts in air traffic control conflict alerting system: Is there a "cry wolf" effect?" *Human factors*, vol. 51, no. 4, pp. 446–462, Aug. 2009.

[74] N. S. Stanton, S. A. Ragsdale, and E. A. Bustamante, "The Effects of System Technology and Probability Type on Trust, Compliance, and Reliance," in *Proceedings of the Human Factors and Ergonomics Society 53rd Annual Meeting*, Jun. 2009, pp. 1368–1372.

[75] R. B. Davenport and E. A. Bustamante, "Effects of False-Alarm vs. Miss-Prone Automation and Likelihood Alarm Technology on Trust, Reliance, and Compliance in a Miss-Prone Task," in *Proceedings of the Human Factors and Ergonomics Society 54th Annual Meeting*, 2010, pp. 1513–1517.

[76] P. Madhavan, D. A. Wiegmann, and F. C. Lacson, "Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids," *Human Factors*, vol. 48, no. 2, pp. 241–256, 2006.

[77] J. Sauer, A. Chavaillaz, and D. Wastell, "Experience of automation failures in training : Effects on trust , automation bias , complacency and performance," *Ergonomics*, vol. 59, no. 6, pp. 767–780, Jun. 2016.

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

[79] A. Chaudhuri and L. Gangadharan, "Gender differences in trust and reciprocity," *The University of Auckland, Department of Economics Working Paper Series*, 2003.

[80] J. Berg, J. Dickhaut, and K. McCabe, "Trust, Reciprocity, and Social History," *Games and Economic Behavior*, vol. 10, no. 1, pp. 122–142, 1995.

[81] V. Venkatesh, M. G. Morris, and P. L. Ackerman, "A longitudinal field investigation of gender differences in individual technology adoption decision-making processes," *Organizational Behavior and Human Decision Processes*, vol. 83, no. 1, pp. 33–60, Sep. 2000.

[82] T. Nomura, T. Kanda, T. Suzuki, and K. Kato, "Prediction of Human Behavior in Human–Robot Interaction Using Psychological Scales for Anxiety and Negative Attitudes Toward Robots," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 442–451, 2008.

[83] M. Siegel, C. Breazeal, and M. I. Norton, "Persuasive robotics: The influence of robot gender on human behavior," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2009*, Oct. 2009, pp. 2563–2568.

[84] F.-W. Tung, "Influence of Gender and Age on the Attitudes of Children towards Humanoid Robots," *Human-Computer Interaction*, vol. IV, pp. 637–646, 2011.

[85] A. Feldhütter, C. Gold, A. Hüger, and K. Bengler, "Trust in Automation as a matter of media and experience of automated vehicles." in *Proceedings of the Human Factors and Ergonomics Society 2016 Annual Meeting*, 2016, pp. 2024–2028.

[86] G. Hofstede, *Culture's Consequences: International Differences in Work Related Values*. SAGE Publications, London, 1980.

[87] P. M. Doney, J. P. Cannon, and M. R. Mullen, "Understanding the influence of national culture on the development of trust," *Academy of Management Review*, vol. 23, no. 3, pp. 601–620, 1998.

[88] D. Gefen and T. Heart, "On the Need to Include National Culture as a Central Issue in E-Commerce Trust Beliefs," *Journal of Global Information Management*, vol. 14, no. 4, pp. 1–30, 2006.

[89] S. Rice, K. Kraemer, S. R. Winter, R. Mehta, V. Dunbar, T. G. Rosser, and J. C. Moore, "Passengers from India and the United States have differential opinions about autonomous auto-pilots for commercial flights," *International Journal of Aviation, Aeronautics, and Aerospace*, vol. 1, no. 1, p. 3, 2014.

[90] E. Huerta, T. Glandon, and Y. Petrides, "Framing, decision-aid systems, and culture: Exploring influences on fraud investigations," *International Journal of Accounting Information Systems*, vol. 13, no. 4, pp. 316–333, 2012.

[91] F. D. Schoorman, R. C. Mayer, and J. H. Davis, "An Integrative Model of Organizational Trust: Past, Present, and Future," *Academy of Management*, vol. 32, no. 2, pp. 344–354, 2007.

[92] G. Hofstede, G. J. Hofstede, and M. Minkov, *Cultures and Organizations: Software of the Mind*, 3rd ed. McGraw-Hill Education, 2010.

[93] C. M. N. Faisal, M. Gonzalez-Rodriguez, D. Fernandez-Lanvin, and J. de Andres-Suarez, "Web Design Attributes in Building User Trust, Satisfaction, and Loyalty for a High Uncertainty Avoidance Culture," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 847–859, Dec. 2017.

[94] A. Vance, C. Elie-Dit-Cosaque, and D. W. Straub, "Examining Trust in Information Technology Artifacts: The Effects of System Quality and Culture," *Journal of Management Information Systems*, vol. 24, no. 4, pp. 73–100, 2008.

[95] I. P. L. Png, B. C. Y. Tan, and K. L. Wee, "Dimensions of national culture and corporate adoption of IT infrastructure," *IEEE Transactions on Engineering Management*, vol. 48, no. 1, pp. 36–45, 2001.

[96] S.-Y. Chien, M. Lewis, K. Sycara, Jyi-Shane Liu, and A. Kumru, "Influence of cultural factors in dynamic trust in automation," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. Budapest, Hungary Influence: IEEE, Oct. 2016, pp. 002 884–002 889.

[97] Amazon, "Amazon Mechanical Turk," *Amazon Mechanical Turk - Welcome*, 2005.

[98] P. J. Rousseeuw and M. Hubert, "Robust statistics for outlier detection," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 1, pp. 73–79, 2011.

[99] D. Pfeffermann, "The Role of Sampling Weights When Modeling Survey Data," *International Statistical Review*, vol. 61, no. 2, pp. 317–337, Aug. 1993.

[100] S. J. Orfanidis, *Introduction to Signal Processing*, ser. Prentice Hall International Editions. Prentice-Hall, Inc., 1995.

[101] M. P. Haselhuhn, J. A. Kennedy, L. J. Kray, A. B. Van Zant, and M. E. Schweitzer, "Gender differences in trust dynamics: Women trust more than men following a trust violation," *Journal of Experimental Social Psychology*, vol. 56, pp. 104–109, 2015.

[102] K. Akash, W.-L. Hu, N. Jain, and T. Reid, "A Classification Model for Sensing Human Trust in Machines Using EEG and GSR," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 4, pp. 1–20, Nov. 2018.

[103] W. D. Penny, S. J. Roberts, E. A. Curran, and M. J. Stokes, "EEG-based communication: A pattern recognition approach," *IEEE Transactions on Rehabilitation Engineering*, vol. 8, no. 2, pp. 214–215, 2000.

[104] G. Pfurtscheller, D. Flotzinger, and J. Kalcher, "Brain-Computer Interface—a new communication device for handicapped persons," *Journal of Microcomputer Applications*, vol. 16, no. 3, pp. 293–299, 1993.

[105] D. Mcfarland, C. Anderson, K.-R. Muller, A. Schlogl, and D. Krusienski, "BCI Meeting 2005—Workshop on BCI Signal Processing: Feature Extraction and Translation," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 14, no. 2, pp. 135–138, Jun. 2006.

[106] Q. Ma, L. Meng, and Q. Shen, "You have my word: Reciprocity expectation modulates feedback-related negativity in the trust game," *PloS one*, vol. 10, no. 2, 2015.

[107] R. Nikula, "Psychological Correlates of Nonspecific Skin Conductance Responses," *Psychophysiology*, vol. 28, no. 1, pp. 86–90, 1991.

[108] S. C. Jacobs, R. Friedman, J. D. Parker, G. H. Tofler, A. H. Jimenez, J. E. Muller, H. Benson, and P. H. Stone, "Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research," *American Heart Journal*, vol. 128, no. 6, pp. 1170–1177, 1994.

[109] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven, "EEG Correlates of Task Engagement and Mental Workload in Vigilance, Learning, and Memory Tasks," *Aviation, Space, and Environmental Medicine*, vol. 78, no. 5, pp. B231–B244, 2007.

[110] H. Blinchikoff and H. Krause, *Filtering in the Time and Frequency Domains*, ser. Classic Series. Noble Publishing, 1976.

[111] W. Levy, "Effect of epoch length on power spectrum analysis of the EEG," *Anesthesiology*, vol. 66, no. 4, pp. 489–495, Apr. 1987.

[112] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain–computer interfaces," *J. Neural Eng.*, vol. 4, no. 2, pp. R1–R13, Jun. 2007.

[113] H. U. Amin, A. S. Malik, R. F. Ahmad, N. Badruddin, N. Kamel, M. Hussain, and W.-T. Chooi, "Feature extraction and classification for EEG signals using wavelet transform and machine learning techniques," *Australasian Physical & Engineering Sciences in Medicine*, vol. 38, no. 1, pp. 139–149, 2015.

[114] D. Sundararajan, *Discrete Wavelet Transform: A Signal Processing Approach*, ser. CourseSmart Series. Wiley, 2016.

[115] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of Neuroscience Methods*, vol. 190, no. 1, pp. 80–91, 2010.

[116] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, no. 1, pp. 273–324, 1997.

[117] I. Kononenko, E. S. Imec, and M. R.-S. Ikonja, "Overcoming the Myopia of Inductive Learning Algorithms with RELIEFF," *Applied Intelligence*, p. 17, 1997.

[118] K. Kira and L. A. Rendell, "A practical approach to feature selection," in *Proceedings of the Ninth International Workshop on Machine Learning*, 1992, pp. 249–256.

[119] P. Pudil, J. Novovičov á, and J. Kittler, "Floating search methods in feature selection," *Pattern Recognition Letters*, vol. 15, no. 11, pp. 1119–1125, 1994.

[120] W.-L. Hu, K. Akash, N. Jain, and T. Reid, "Real-Time Sensing of Trust in Human-Machine Interactions," in *1st IFAC Conference on Cyber-Physical & Human-Systems*, Florianopolis, Brazil, 2016.

[121] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, ser. Springer Series in Statistics. Springer New York, 2009.

[122] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median," *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.

[123] N. Jaušovec and K. Jaušovec, "EEG activity during the performance of complex mental problems," *International Journal of Psychophysiology*, vol. 36, no. 1, pp. 73–88, 2000.

[124] S. Righi, L. Mecacci, and M. P. Viggiano, "Anxiety, cognitive self–evaluation and performance: ERP correlates," *Journal of Anxiety Disorders*, vol. 23, no. 8, pp. 1132–1138, 2009.

[125] C. Dussault, J.-C. Jouanin, M. Philippe, and C.-Y. Guezennec, "EEG and ECG Changes During Simulator Operation Reflect Mental Workload and Vigilance," *Aviation, Space, and Environmental Medicine*, vol. 76, no. 4, 2005.

[126] W. J. Ray and H. W. Cole, "EEG alpha activity reflects attentional demands, and beta activity reflects emotional and cognitive processes." *Science*, vol. 228, no. 4700, pp. 750–752, 1985.

[127] T. Isotani, H. Tanaka, D. Lehmann, R. D. Pascual-Marqui, K. Kochi, N. Saito, T. Yagyu, T. Kinoshita, and K. Sasada, "Source localization of EEG activity during hypnotically induced anxiety and relaxation," *International Journal of Psychophysiology*, vol. 41, no. 2, pp. 143–153, 2001.

[128] F. Chen, N. Ruiz, E. Choi, J. Epps, M. A. Khawaja, R. Taib, B. Yin, and Y. Wang, "Multimodal behavior and interaction as indicators of cognitive load," *ACM Transactions on Interactive Intelligent Systems*, vol. 2, no. 4, pp. 1–36, Dec. 2012.

[129] J. Zhou, J. Sun, F. Chen, Y. Wang, R. Taib, A. Khawaji, and Z. Li, "Measurable Decision Making with GSR and Pupillary Analysis for Intelligent User Interface," *ACM Transactions on Computer-Human Interaction*, vol. 21, no. 6, pp. 33:1–33:23, Jan. 2015.

[130] Mathworks, "Statistics and Machine Learning Toolbox: User's Guide (r2016b)," 2016.

[131] R. Riedl, M. Hubert, and P. Kenning, "Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers," *Management Information Systems Quarterly*, vol. 34, no. 2, pp. 397–428, 2010.

[132] K. Akash, T. Reid, and N. Jain, "Adaptive Probabilistic Classification of Dynamic Processes: A Case Study on Human Trust in Automation," in *2018 Annual American Control Conference (ACC)*. IEEE, Jun. 2018, pp. 246–251.

[133] S. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques," *Informatica*, vol. 31, pp. 249–268, 2007.

[134] D. Tan and A. Nijholt, "Brain-Computer Interfaces and Human-Computer Interaction," in *Brain-Computer Interfaces: Applying Our Minds to Human-Computer Interaction*, D. S. Tan and A. Nijholt, Eds. London: Springer London, 2010, pp. 3–19.

[135] B. Obermaier, C. Guger, C. Neuper, and G. Pfurtscheller, "Hidden Markov models for online classification of single trial EEG data," *Pattern recognition letters*, vol. 22, no. 12, pp. 1299–1309, 2001.

[136] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface," *IEEE Transactions on neural systems and rehabilitation engineering*, vol. 9, no. 3, pp. 283–288, 2001.

[137] F. Cincotti, A. Scipione, A. Timperi, D. Mattia, A. Marciani, J. Millan, S. Salinari, L. Bianchi, and F. Bablioni, "Comparison of different feature classifiers for brain computer interfaces," in *Neural Engineering, 2003. Conference Proceedings. First International IEEE EMBS Conference On.* IEEE, 2003, pp. 645–647.

[138] A. Y. Ng and M. I. Jordan, "On Discriminative vs. Generative Classifiers: A comparison of logistic regression and naive Bayes," in *Advances in Neural Information Processing Systems 14*, T. G. Dietterich, S. Becker, and Z. Ghahramani, Eds. MIT Press, 2002, pp. 841–848.

[139] G. D. Forney, "The Viterbi algorithm," *Proceedings of the IEEE*, vol. 61, no. 3, pp. 268–278, 1973.

[140] C. Anagnostopoulos, D. K. Tasoulis, N. M. Adams, N. G. Pavlidis, and D. J. Hand, "Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification," *Statistical Analysis and Data Mining*, vol. 5, no. 2, pp. 139–166, Apr. 2012.

[141] S. Theodoridis and K. Koutroumbas, *Pattern Recognition*, ser. Pattern Recognition Series. Elsevier Science, 2006.

[142] K. Akash, G. McMahon, T. Reid, and N. Jain, "Human Trust-based Feedback Control: Dynamically varying automation transparency to optimize human-machine interactions," *IEEE Control Systems Magazine*, pp. 1–16. (Accepted), 2020.

[143] N. Moray and T. Inagaki, "Laboratory studies of trust between humans and machines in automated systems," *Transactions of the Institute of Measurement and Control*, vol. 21, no. 4-5, pp. 203–211, Oct. 1999.

[144] B. M. Muir, "Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems," *Ergonomics*, vol. 37, no. 11, pp. 1905–1922, 1994.

[145] J.-Y. Jian, A. M. Bisantz, and C. G. Drury, "Foundations for an Empirically Determined Scale of Trust in Automated Systems," *International Journal of Cognitive Ergonomics*, vol. 4, no. 1, pp. 53–71, 2000.

[146] M. Desai, "Modeling Trust to Improve Human-Robot Interaction," Ph.D., University of Massachusetts Lowell, United States – Massachusetts, 2012.

[147] P. de Vries, C. Midden, and D. Bouwhuis, "The effects of errors on system trust, self-confidence, and the allocation of control in route planning," *International Journal of Human-Computer Studies*, vol. 58, no. 6, pp. 719–735, Jun. 2003.

[148] B. M. Muir and N. Moray, "Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation," *Ergonomics*, vol. 39, no. 3, pp. 429–460, 1996.

[149] N. Moray, T. Inagaki, and M. Itoh, "Adaptive automation, trust, and self-confidence in fault management of time-critical tasks," *Journal of Experimental Psychology: Applied*, vol. 6, no. 1, pp. 44–58, 2000.

[150] a. J. D. Lee, "Extending the decision field theory to model operators' reliance on automation in supervisory control situations," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 36, no. 5, pp. 943–959, Sep. 2006.

[151] M. S. Cohen, R. Parasuraman, and J. T. Freeman, "Trust in decision aids: A model and its training implications," in *In Proc. Command and Control Research and Technology Symp*, 1998, pp. 1–37.

[152] A. Xu and G. Dudek, "OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*, ser. HRI '15. New York, NY, USA: ACM, 2015, pp. 221–228.

[153] E. ElSalamouny, V. Sassone, and M. Nielsen, "HMM-based trust model," in *International Workshop on Formal Aspects in Security and Trust*. Springer, Berlin, Heidelberg, 2009, pp. 21–35.

[154] M. Li and A. M. Okamura, "Recognition of operator motions for real-time assistance using virtual fixtures," in *Haptic Interfaces for Virtual Environment and Teleoperator Systems, 2003. HAPTICS 2003. Proceedings. 11th Symposium On*. IEEE, 2003, pp. 125–131.

[155] J. Pineau, G. Gordon, S. Thrun *et al.*, "Point-based value iteration: An anytime algorithm for POMDPs," in *IJCAI*, vol. 3, 2003, pp. 1025–1032.

[156] Z. Wang, A. Peer, and M. Buss, "An HMM approach to realistic haptic human-robot interaction," in *EuroHaptics Conference, 2009 and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems. World Haptics 2009. Third Joint*. IEEE, 2009, pp. 374–379.

[157] X. Liu and A. Datta, "Modeling Context Aware Dynamic Trust Using Hidden Markov Model," in *AAAI*, 2012, pp. 1938–1944.

[158] L. Feng, C. Wiltsche, L. Humphrey, and U. Topcu, "Synthesis of Human-in-the-Loop Control Protocols for Autonomous Systems," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 450–462, Apr. 2016.

[159] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Planning with Trust for Human-Robot Collaboration," in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*. Chicago, IL, USA: ACM Press, 2018, pp. 307–315.

[160] R. Seymour and G. L. Peterson, "A Trust-Based Multiagent System," in *2009 International Conference on Computational Science and Engineering*, vol. 3, Aug. 2009, pp. 109–116.

[161] R. W. Proctor and T. Van Zandt, *Human Factors in Simple and Complex Systems*, 3rd ed. CRC Press, 2018.

[162] R. D. Luce, *Response Times: Their Role in Inferring Elementary Mental Organization*. OUP USA, Jul. 1986.

[163] J. Braithwaite and T. Makkai, "Trust and compliance," *Policing and Society*, vol. 4, no. 1, pp. 1–12, May 1994.

[164] J. E. Fox, "The Effects of Information Accuracy on User Trust and Compliance," in *Conference Companion on Human Factors in Computing Systems*, ser. CHI '96. Vancouver, British Columbia, Canada: ACM Press, 1996, pp. 35–36.

[165] J. E. Fox and D. A. Boehm-Davis, "Effects of Age and Congestion Information Accuracy of Advanced Traveler Information Systems on User Trust and Compliance," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1621, no. 1, pp. 43–49, Jan. 1998.

[166] D. E. Crundall, G. J. Underwood, and P. R. Chapman, "Peripheral detection rates in drivers," *Vision in vehicles*, vol. 7, pp. 261–269, 1999.

[167] C. J. Patten, A. Kircher, J. Östlund, and L. Nilsson, "Using mobile telephones: Cognitive workload and attention resource allocation," *Accident Analysis & Prevention*, vol. 36, no. 3, pp. 341–350, May 2004.

[168] G. S. Newell and N. J. Mansfield, "Evaluation of reaction time performance and subjective workload during whole-body vibration exposure while seated in upright and twisted postures with and without armrests," *International Journal of Industrial Ergonomics*, vol. 38, no. 5-6, pp. 499–508, May 2008.

[169] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[170] O. Sigaud and O. Buffet, *Markov Decision Processes in Artificial Intelligence*. John Wiley & Sons, Mar. 2013.

[171] D. A. Balota and D. H. Spieler, "Word frequency, repetition, and lexicality effects in word recognition tasks: Beyond measures of central tendency." *Journal of experimental psychology. General*, vol. 128, no. 1, pp. 32–55, 1999.

[172] R. Whelan, "Effective Analysis of Reaction Time Data," *The Psychological Record*, vol. 58, no. 3, pp. 475–482, Jul. 2008.

[173] P. Jaśkowski, "Distribution of the human reaction time measurements," *Acta Neurobiol Exp (Wars)*, vol. 43, no. 3, pp. 221–225, 1983.

[174] R. Ratcliff and B. B. Murdock, "Retrieval processes in recognition memory," *Psychological Review*, vol. 83, no. 3, pp. 190–214, 1976.

[175] W. J. McGill and J. Gibbon, "The general-gamma distribution and reaction times," *Journal of Mathematical Psychology*, vol. 2, no. 1, pp. 1–18, Feb. 1965.

[176] Y. Lacouture and D. Cousineau, "How to use MATLAB to fit the ex-Gaussian and other probability functions to a distribution of response times," *Tutorials in Quantitative Methods for Psychology*, vol. 4, no. 1, pp. 35–45, Mar. 2008.

[177] B. A. Parris, Z. Dienes, and T. L. Hodgson, "Application of the ex-Gaussian function to the effect of the word blindness suggestion on Stroop task performance suggests no word blindness," *Front. Psychol.*, vol. 4, 2013.

[178] C. Moret-Tatay, D. Gamermann, E. Navarro-Pardo, and P. Fernández de Córdoba Castellá, "ExGUtils: A Python Package for Statistical Analysis With the ex-Gaussian Probability Density," *Front. Psychol.*, vol. 9, 2018.

[179] A. Heathcote, S. J. Popiel, and D. J. Mewhort, "Analysis of response time distributions: An example using the Stroop task." *Psychological Bulletin*, vol. 109, no. 2, pp. 340–347, 1991.

[180] E. Navarro-Pardo, A. B. Navarro-Prados, D. Gamermann, and C. Moret-Tatay, "Differences Between Young and Old University Students on a Lexical Decision Task: Evidence Through an Ex-Gaussian Approach," *The Journal of General Psychology*, vol. 140, no. 4, pp. 251–268, Oct. 2013.

[181] D. Gooch, M. J. Snowling, and C. Hulme, "Reaction Time Variability in Children With ADHD Symptoms and/or Dyslexia," *Developmental Neuropsychology*, vol. 37, no. 5, pp. 453–472, Jul. 2012.

[182] A. S. Hervey, J. N. Epstein, J. F. Curry, S. Tonev, L. Eugene Arnold, C. Keith Conners, S. P. Hinshaw, J. M. Swanson, and L. Hechtman, "Reaction Time Distribution Analysis of Neuropsychological Performance in an ADHD Sample," *Child Neuropsychology*, vol. 12, no. 2, pp. 125–140, May 2006.

[183] R. H. Hohle, "Inferred components of reaction times as functions of foreperiod duration," *Journal of Experimental Psychology*, vol. 69, no. 4, pp. 382–386, Apr. 1965.

[184] D. Matzke and E.-J. Wagenmakers, "Psychological interpretation of the ex-Gaussian and shifted Wald parameters: A diffusion model analysis," *Psychonomic Bulletin & Review*, vol. 16, no. 5, pp. 798–817, Oct. 2009.

[185] J. J. Shaughnessy, E. B. Zechmeister, and J. S. Zechmeister, *Research Methods in Psychology*, 9th ed. New York, NY: McGraw-Hill, 2012.

[186] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, Jan. 1986.

[187] S. M. Merritt and D. R. Ilgen, "Not all trust is created equal: Dispositional and history-based trust in human-automation interactions," *Human Factors*, vol. 50, no. 2, pp. 194–210, 2008.

[188] A. R. Cassandra, L. P. Kaelbling, and M. L. Littman, "Acting optimally in partially observable stochastic domains," in *AAAI*, vol. 94, 1994, pp. 1023–1028.

[189] R Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2018.

[190] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting Linear Mixed-Effects Models Using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[191] N. A. Stanton and M. S. Young, "Vehicle automation and driving performance," *Ergonomics*, vol. 41, no. 7, pp. 1014–1028, Jul. 1998.

[192] R. D. Spain and J. P. Bliss, "The effect of sonification display pulse rate and reliability on operator trust and perceived workload during a simulated patient monitoring task," *Ergonomics*, vol. 51, no. 9, pp. 1320–1337, Sep. 2008.

[193] T. Sato, Y. Yamani, M. Liechty, and E. T. Chancey, "Automation trust increases under high-workload multitasking scenarios involving risk," *Cogn Tech Work*, vol. 22, no. 2, pp. 399–407, May 2020.

[194] N. D. Karpinsky, E. T. Chancey, D. B. Palmer, and Y. Yamani, "Automation trust and attention allocation in multitasking workspace," *Applied Ergonomics*, vol. 70, pp. 194–201, Jul. 2018.

A. TRUST AND WORKLOAD POMDP MODELS

We present the estimated independent, coupled-transition, and coupled-emission POMDP models of human trust-workload behavior discussed in Chapter 4.

A.1 Independent Model

The independent model for trust and workload behavior represented in Figure 4.1 consists of two independent POMDP models: a trust model and a workload model.

Trust Model

The estimated trust model consists of initial state probabilities $\pi(s_T)$, an emission probability function $\mathcal{E}_T(o_C|s_T)$, and a transition probability function $\mathcal{T}_T(s'_T|s_T, a)$. Based on the emission probability function for trust $\mathcal{E}_T(o_C|s_T)$, we define the High Trust state $s_T = T_\uparrow$ as that in which there is a higher probability of observing the human comply with the automation's recommendation, $o_C = C^+$. The estimated initial probabilities of Low Trust $T_\downarrow$ and High Trust $T_\uparrow$ are $\pi(T_\downarrow) = 0.1283$ and $\pi(T_\uparrow) = 0.8717$, respectively. The emission probability function $\mathcal{E}_T(o_C|s_T)$ is depicted in Figure A.1 and characterizes the probability of a participant's compliance with the system's recommendations given the participant's state of trust.

Figure A.2 represents the transition probability function $\mathcal{T}_T(s'_T|s_T, a)$ showing the probability of transitioning from the state s_T to s'_T (where $s_T, s'_T \in T$) given the action $a \in \mathcal{A}$.

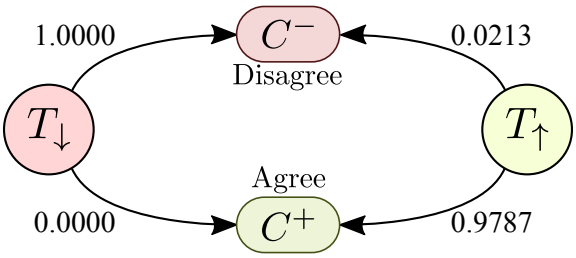


Figure A.1. Emission probability function $\mathcal{E}_T(o_C|s_T)$ for trust in the independent model. Probabilities of observation are shown beside the arrows.

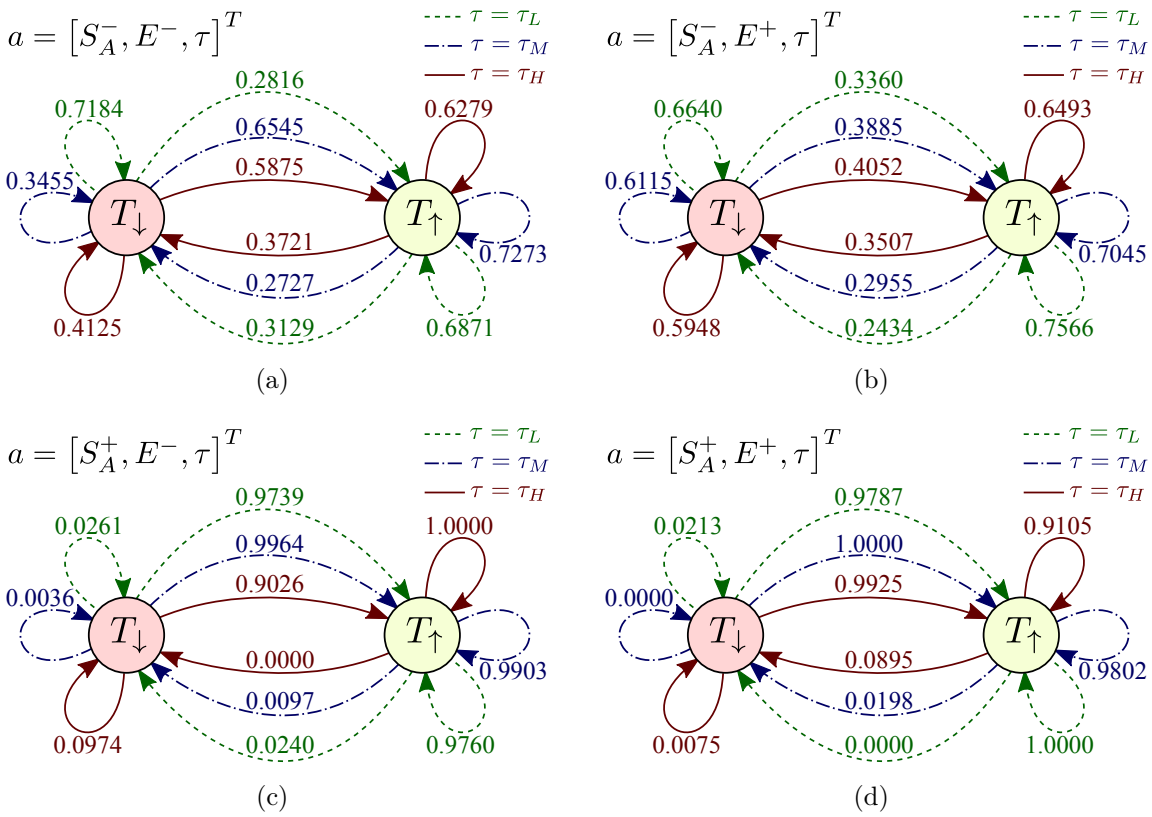


Figure A.2. Transition probability function $\mathcal{T}_T(s'_T|s_T, a)$ for trust in the independent model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid's recommendation is Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

Workload Model

The workload model consists of initial state probabilities $\pi(s_W)$, an emission probability function $\mathcal{E}_W(o_{RT}|s_W)$, and a transition probability function $\mathcal{T}_W(s'_W|s_W, a)$. Similar to the trust model, based on the emission probability function for workload $\mathcal{E}_W(o_{RT}|s_W)$, we define the High Workload state $s_W = W_\uparrow$ as that in which the expected response time $\mathbf{E}[o_{RT}|s_W]$ is longer. We estimated the initial probabilities of Low Workload $W_\downarrow$ and High Workload $W_\uparrow$ to be $\pi(W_\downarrow) = 0.3487$ and $\pi(W_\uparrow) = 0.6513$, respectively. The emission probability function $\mathcal{E}_W(o_{RT}|s_W)$ is represented in Figure A.3, which shows the probability density functions (PDFs) of observing participants' response time as o_{RT} given their state of workload s_W.

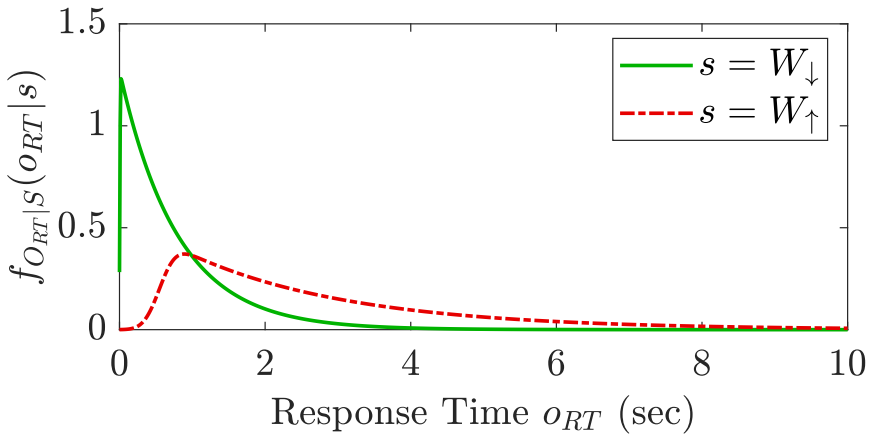


Figure A.3. Emission probability function $\mathcal{E}_W(o_{RT}|s_W)$ for workload in the independent model. For Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_\downarrow}(o_{RT}|W_\downarrow)$ is characterized by an ex-Gaussian distribution with $\mu_{W_\downarrow} = 0.0047$, $\sigma_{W_\downarrow} = 0.0062$, and $\tau_{W_\downarrow} = 0.7917$. For High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_\uparrow}(o_{RT}|W_\uparrow)$ is characterized by an ex-Gaussian distribution with $\mu_{W_\uparrow} = 0.5581$, $\sigma_{W_\uparrow} = 0.1745$, and $\tau_{W_\uparrow} = 2.2544$.

The transition probability function $\mathcal{T}_W(s'_W|s_W, a)$ is represented in Figure A.4 and shows the probability of a participant transitioning from the state s_W to s'_W based on the action $a \in \mathcal{A}$, where $s_W, s'_W \in W$.

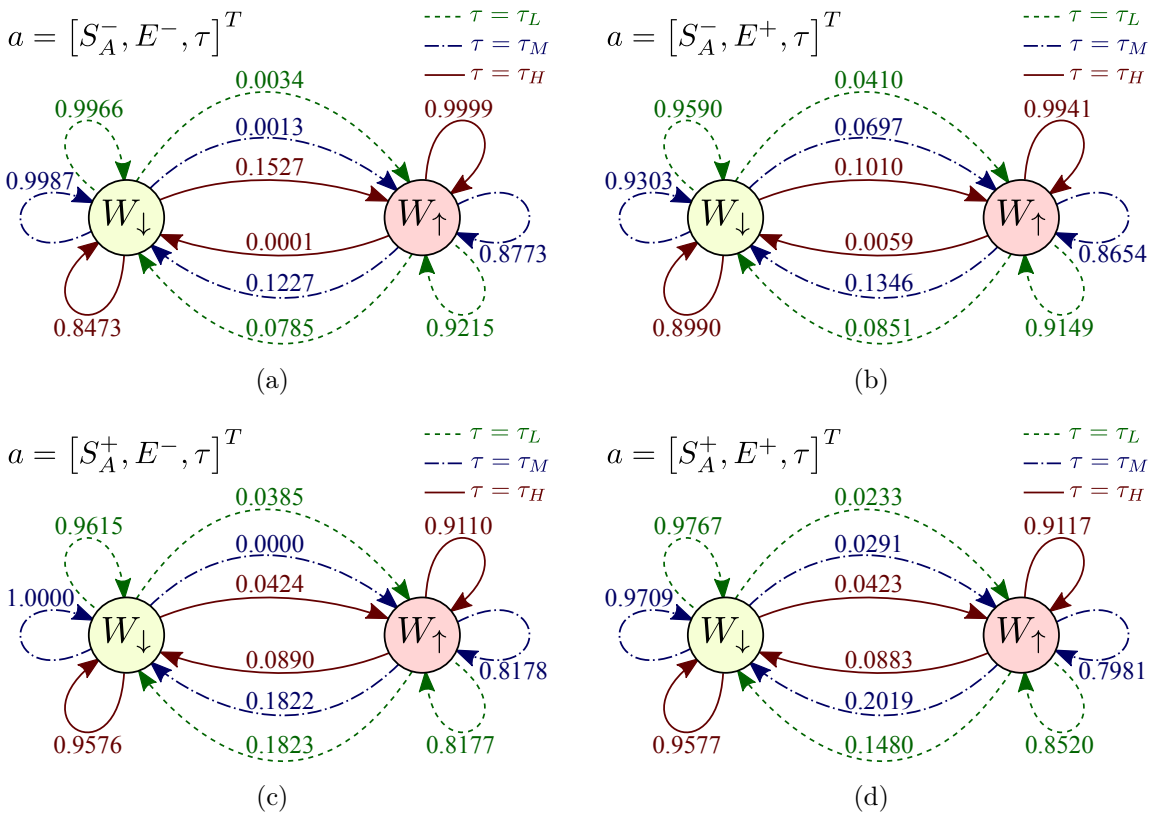


Figure A.4. Transition probability function $\mathcal{T}_W(s'_W|s_W, a)$ for workload in the independent model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

A.2 Coupled-Transition Model

The coupled-transition model for trust and workload behavior represented in Figure 4.2 consists of two coupled POMDP models: a trust model and a workload model, which interact in their transition probabilities.

Trust Model

The estimated trust model consists of initial state probabilities $\pi(s_T)$, an emission probability function $\mathcal{E}_T(o_C|s_T)$, and a transition probability function $\mathcal{T}_T(s'_T|s_T, s_W, a)$. Based on the emission probability function for trust $\mathcal{E}_T(o_C|s_T)$, we define the High Trust state $s_T = T_\uparrow$ as that in which there is a higher probability of observing the human comply with the automation's recommendation, $o_C = C^+$. The estimated initial probabilities of Low Trust $T_\downarrow$ and High Trust $T_\uparrow$ are $\pi(T_\downarrow) = 0.1662$ and $\pi(T_\uparrow) = 0.8338$, respectively. The emission probability function $\mathcal{E}_T(o_C|s_T)$ is depicted in Figure A.5 and characterizes the probability of a participant's compliance with the system's recommendations given the participant's state of trust.

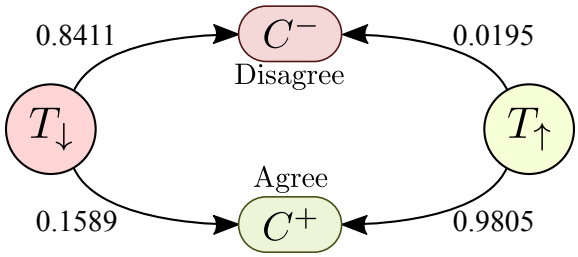


Figure A.5. Emission probability function $\mathcal{E}_T(o_C|s_T)$ for trust in the coupled-transition model. Probabilities of observation are shown beside the arrows.

Figure A.6 represents the transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\downarrow, a)$ showing the probability of transitioning from the state s_T to s'_T (where $s_T, s'_T \in T$) given the workload state is $W_\downarrow$ and the action $a \in \mathcal{A}$.

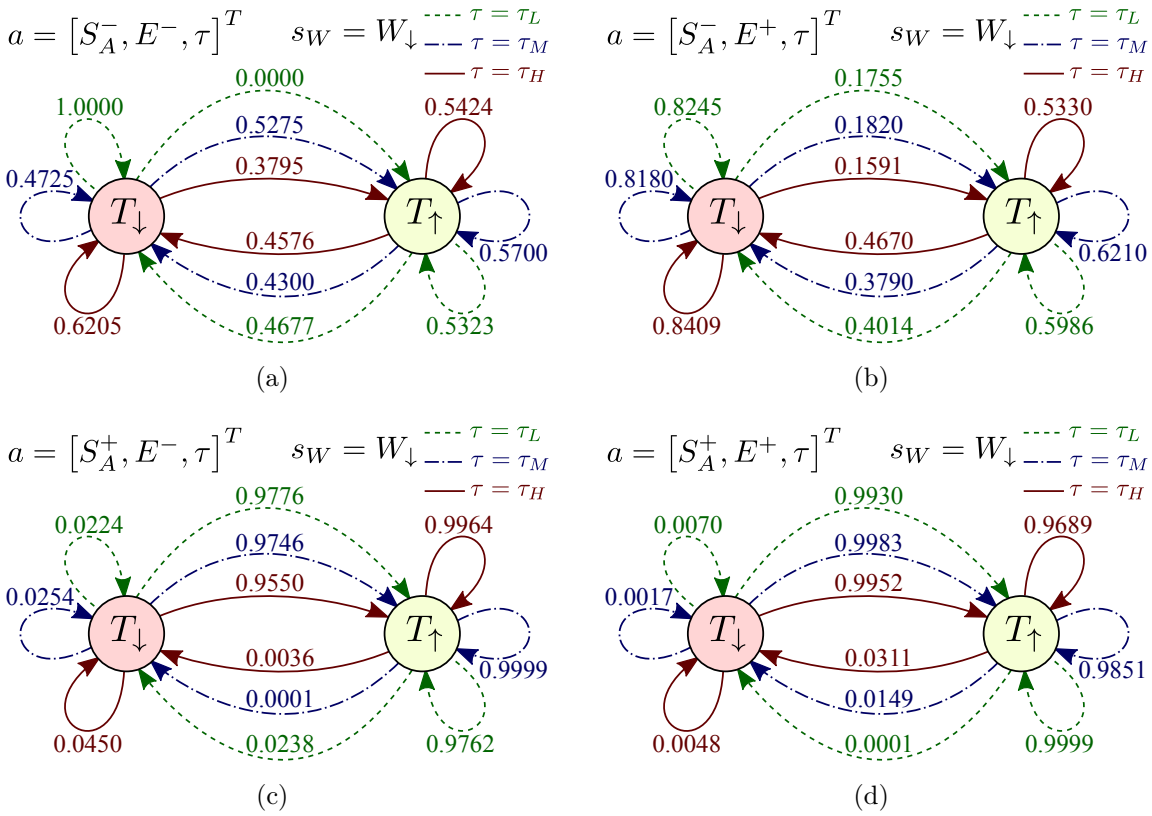


Figure A.6. Transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\downarrow, a)$ for trust in the coupled-transition model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid's recommendation is Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

Figure A.7 represents the transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\uparrow, a)$ showing the probability of transitioning from the state s_T to s'_T (where $s_T, s'_T \in T$) given the workload state is $W_\uparrow$ and the action $a \in \mathcal{A}$.

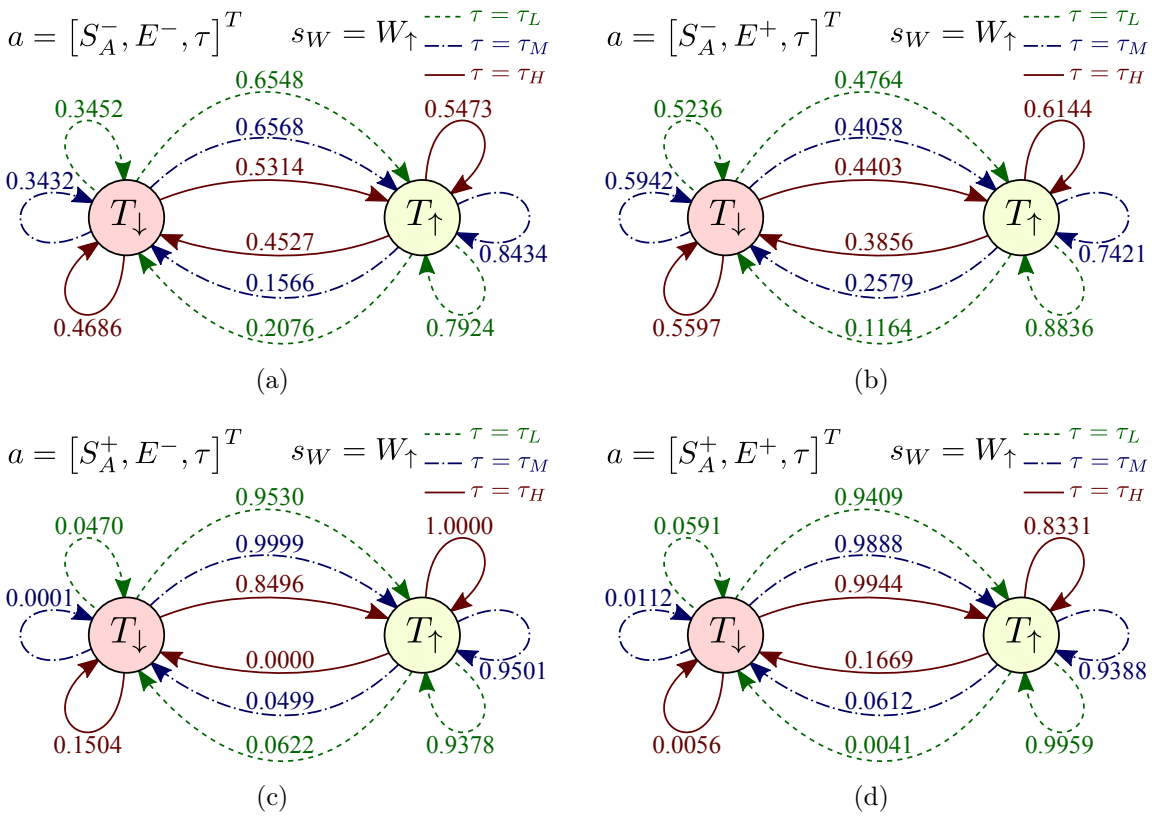


Figure A.7. Transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\uparrow, a)$ for trust in the coupled-transition model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid's recommendation is Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

Workload Model

The workload model consists of initial state probabilities $\pi(s_W)$, an emission probability function $\mathcal{E}_W(o_{RT}|s_W)$, and a transition probability function $\mathcal{T}_W(s'_W|s_T, s_W, a)$. Similar to the trust model, based on the emission probability function for workload $\mathcal{E}_W(o_{RT}|s_W)$, we define the High Workload state $s_W = W_\uparrow$ as that in which the expected response time $\mathbf{E}[o_{RT}|s_W]$ is longer. We estimated the initial probabilities of Low Workload $W_\downarrow$ and High Workload $W_\uparrow$ to be $\pi(W_\downarrow) = 0.3342$ and $\pi(W_\uparrow) = 0.6658$, respectively. The emission probability function $\mathcal{E}_W(o_{RT}|s_W)$ is represented in Figure A.8, which shows the probability density functions (PDFs) of observing participants' response time as o_{RT} given their state of workload s_W.

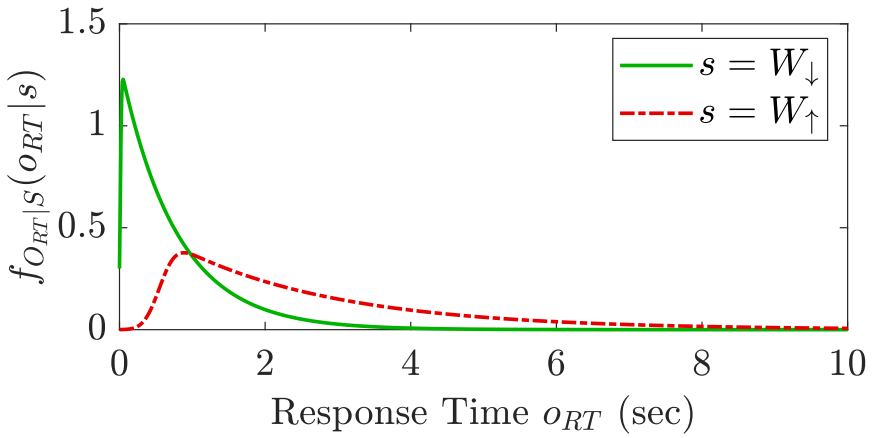


Figure A.8. Emission probability function $\mathcal{E}_W(o_{RT}|s_W)$ for workload in the coupled-transition model. For Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_\downarrow}(o_{RT}|W_\downarrow)$ is characterized by an ex-Gaussian distribution with $\mu_{W_\downarrow} = 0.0108$, $\sigma_{W_\downarrow} = 0.0149$, and $\tau_{W_\downarrow} = 0.7708$. For High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|W_\uparrow}(o_{RT}|W_\uparrow)$ is characterized by an ex-Gaussian distribution with $\mu_{W_\uparrow} = 0.5566$, $\sigma_{W_\uparrow} = 0.1717$, and $\tau_{W_\uparrow} = 2.2179$.

The transition probability function $\mathcal{T}_W(s'_W|s_T = T_\downarrow, s_W, a)$ is represented in Figure A.9 and shows the probability of a participant transitioning from the state s_W to s'_W given the trust state $T_\downarrow$ and the action $a \in \mathcal{A}$.

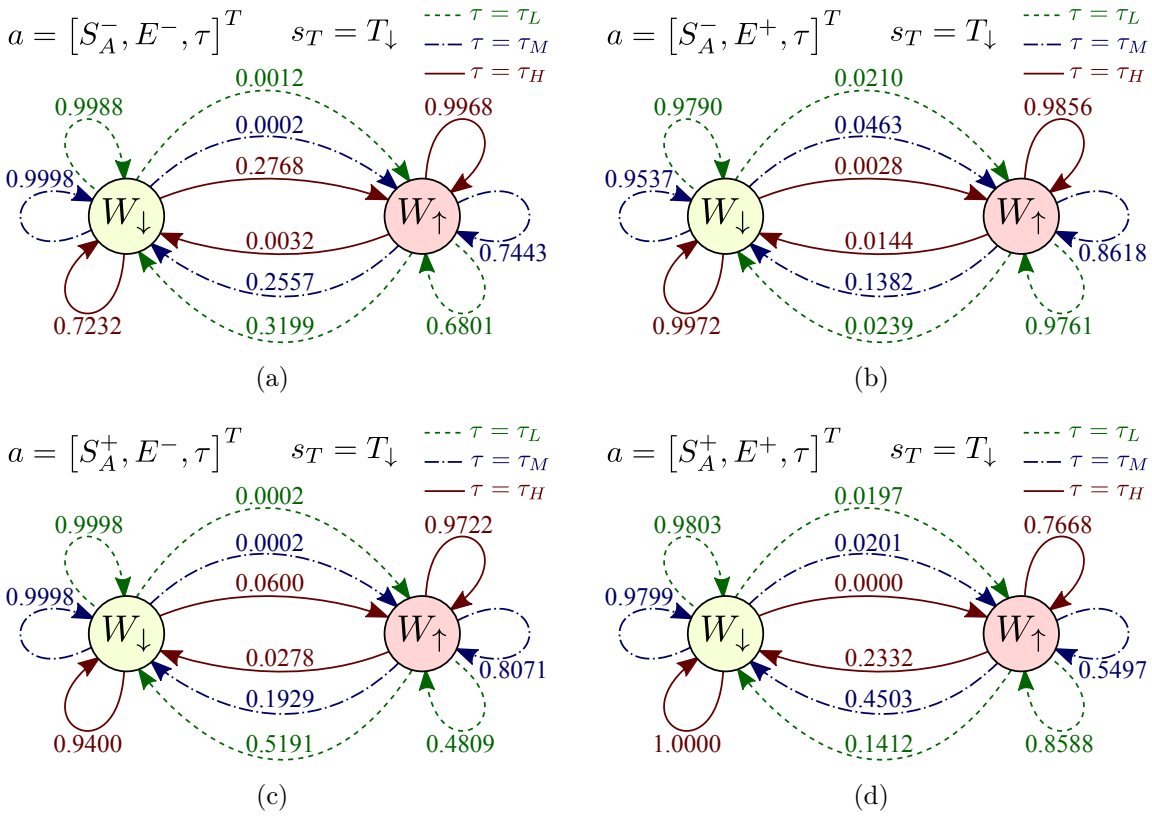


Figure A.9. Transition probability function $\mathcal{T}_W(s'_W|s_T = T_\downarrow, s_W, a)$ for workload in the coupled-transition model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

The transition probability function $\mathcal{T}_W(s'_W|s_T = T_\uparrow, s_W, a)$ is represented in Figure A.10 and shows the probability of a participant transitioning from the state s_W to s'_W given the trust state $T_\uparrow$ and the action $a \in \mathcal{A}$.

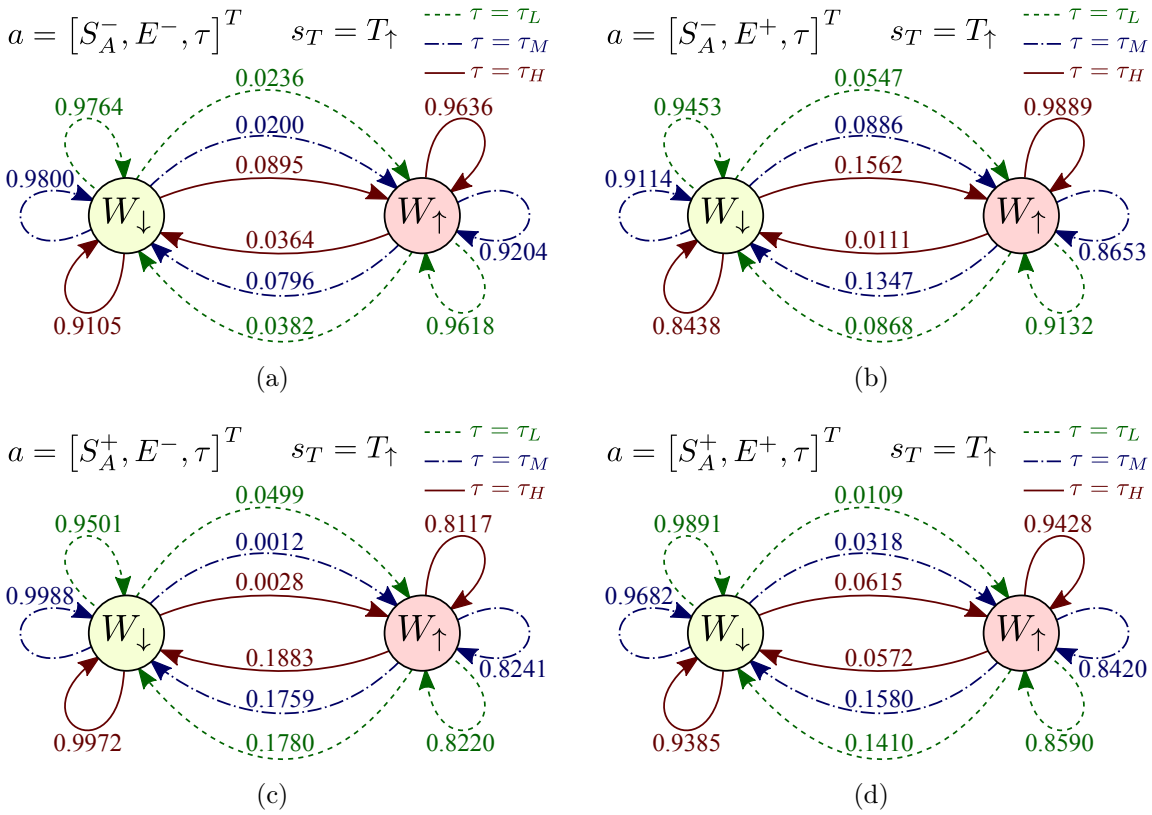


Figure A.10. Transition probability function $\mathcal{T}_W(s'_W|s_T = T_\uparrow, s_W, a)$ for workload in the coupled-transition model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

A.3 Coupled-Emission Model

The coupled-emission model for trust and workload behavior represented in Figure 4.3 consists of two coupled POMDP models: a trust model and a workload model, which interact in their transition as well as their emission probabilities.

Trust Model

The estimated trust model consists of initial state probabilities $\pi(s_T)$, an emission probability function $\mathcal{E}(o_C|s_T, s_W)$, and a transition probability function $\mathcal{T}(s'_T|s_T, s_W, a)$. In this case, identifying the Low Trust and High Trust states is not trivial using the emission probability function $\mathcal{E}(o_C|s_T, s_W)$. Since the emission probability function is dependent on two sets of states, one corresponding to trust states and the other corresponding to the workload states, we first need to identify each set of states as being the set of trust states or workload states. We define the set of states having a larger variation across the probability of compliance as the set of trust states. This is based on the assumption that trust has a stronger influence on compliance as compared to workload. Then among this set of states, we define the High Trust state $s_T = T_\uparrow$ as that in which there is a higher probability of observing the human comply with the automation's recommendation, $o_C = C^+$. The estimated initial probabilities of Low Trust $T_\downarrow$ and High Trust $T_\uparrow$ are $\pi(T_\downarrow) = 0.4877$ and $\pi(T_\uparrow) = 0.5123$, respectively. The emission probability function $\mathcal{E}_T(o_C|s_T, s_W)$ is depicted in Figure A.11 and characterizes the probability of a participant's compliance with the system's recommendations given the participant's state of trust and workload.

Figure A.12 represents the transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\downarrow, a)$ showing the probability of transitioning from the state s_T to s'_T (where $s_T, s'_T \in T$) given the workload state is $W_\downarrow$ and the action $a \in \mathcal{A}$.

Figure A.13 represents the transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\uparrow, a)$ showing the probability of transitioning from the state s_T to s'_T (where $s_T, s'_T \in T$) given the workload state is $W_\uparrow$ and the action $a \in \mathcal{A}$.

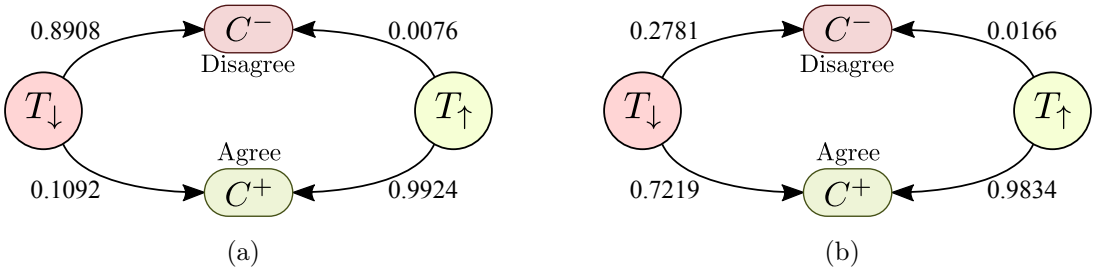


Figure A.11. Emission probability function $\mathcal{E}_T(o_C|s_T, s_W)$ for trust in the coupled-emission model. Probabilities of observation are shown beside the arrows. The left diagram (a) shows the emission probabilities when the workload state is $W_\downarrow$. The right diagram (b) shows the emission probabilities when the workload state is $W_\uparrow$.

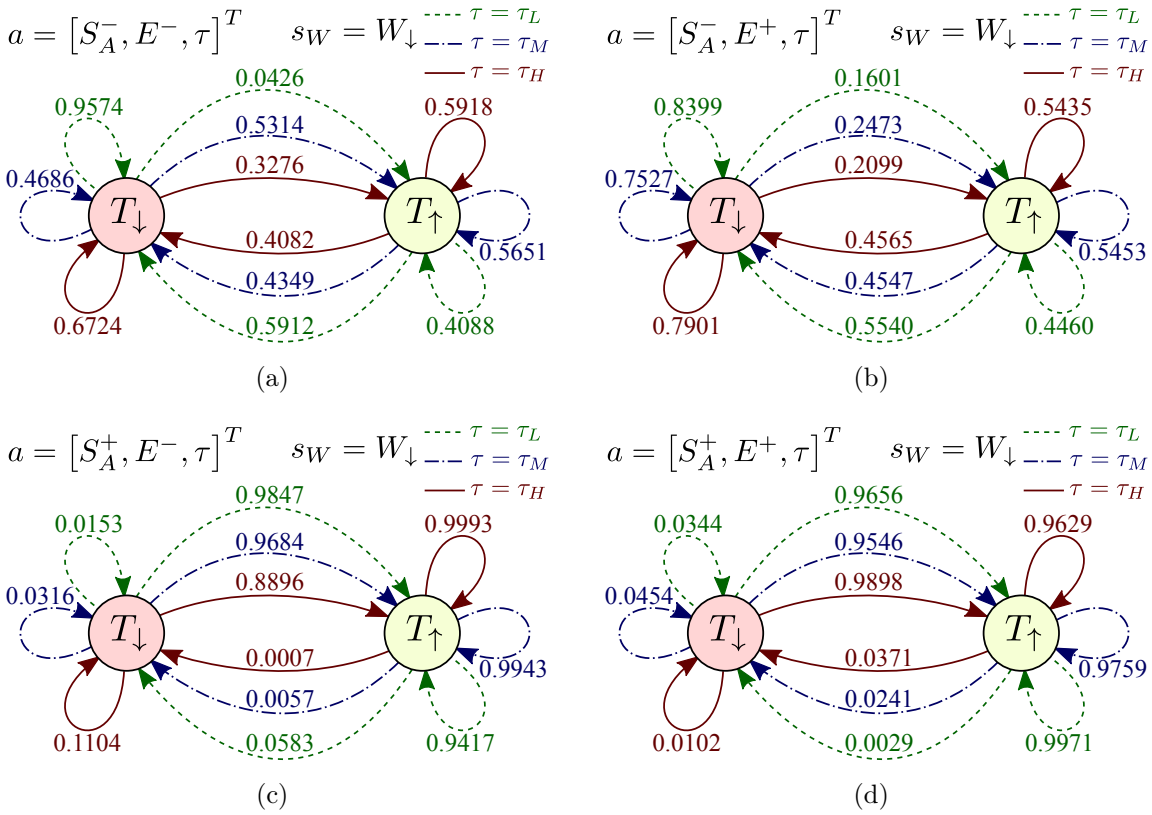


Figure A.12. Transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\downarrow, a)$ for trust in the coupled-emission model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid's recommendation is Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

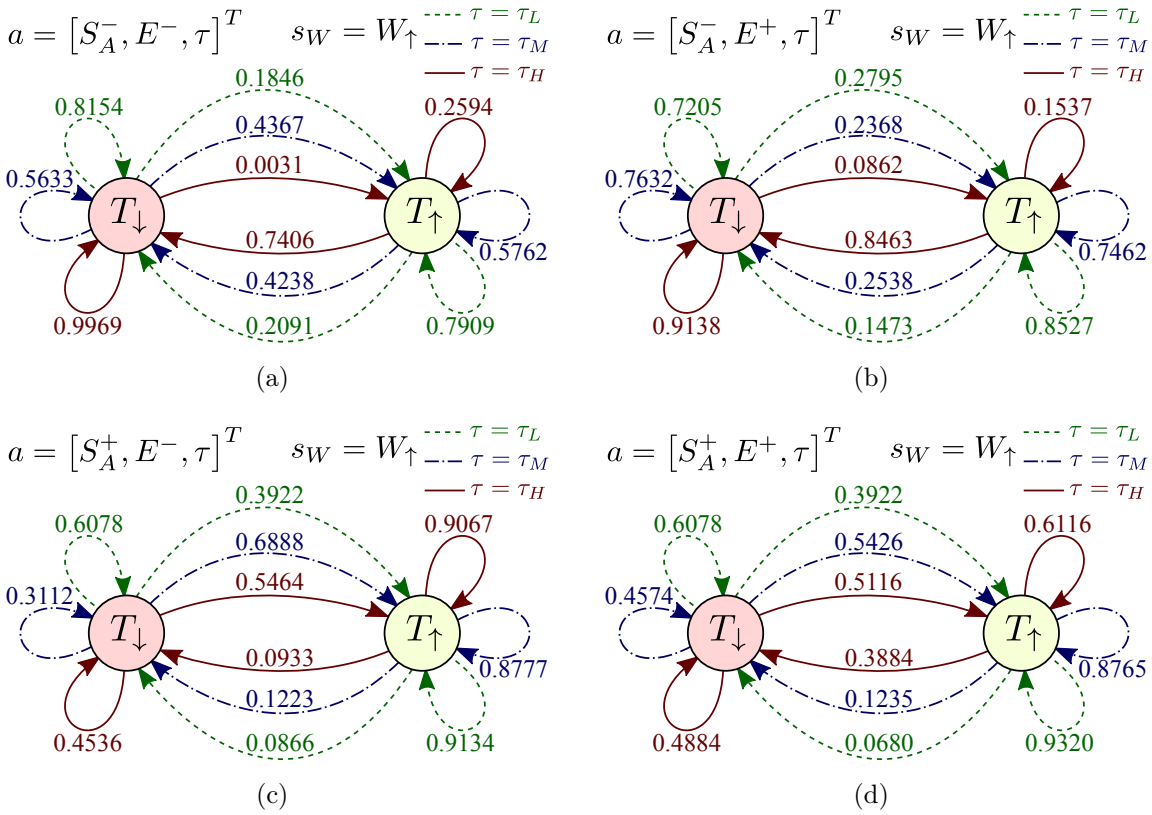


Figure A.13. Transition probability function $\mathcal{T}_T(s'_T|s_T, s_W = W_\uparrow, a)$ for trust in the coupled-emission model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid's recommendation is Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

Workload Model

The workload model consists of initial state probabilities $\pi(s_W)$, an emission probability function $\mathcal{E}(o_{RT}|s_T, s_W)$, and a transition probability function $\mathcal{T}(s'_W|s_T, s_W, a)$. As we have already identified which set of states corresponds to the trust states, the other set of states is therefore the workload states. Then, based on the emission probability function for response time $\mathcal{E}(o_{RT}|s_T, s_W)$, we define the High Workload state $s_W = W_\uparrow$ as that in which the expected response time $\mathbf{E}[o_{RT}|s_W]$ is longer. We estimated the initial probabilities of Low Workload $W_\downarrow$ and High Workload $W_\uparrow$ to be $\pi(W_\downarrow) = 0.2349$ and $\pi(W_\uparrow) = 0.7651$, respectively. The emission probability function $\mathcal{E}_W(o_{RT}|s_T, s_W)$ is represented in Figure A.14, which shows the probability density functions (PDFs) of observing participants' response time as o_{RT} given their state of trust s_T and workload s_W.

The transition probability function $\mathcal{T}_W(s'_W|s_T = T_\downarrow, s_W, a)$ is represented in Figure A.15 and shows the probability of a participant transitioning from the state s_W to s'_W given the trust state $T_\downarrow$ and the action $a \in \mathcal{A}$.

The transition probability function $\mathcal{T}_W(s'_W|s_T = T_\uparrow, s_W, a)$ is represented in Figure A.16 and shows the probability of a participant transitioning from the state s_W to s'_W given the trust state $T_\uparrow$ and the action $a \in \mathcal{A}$.

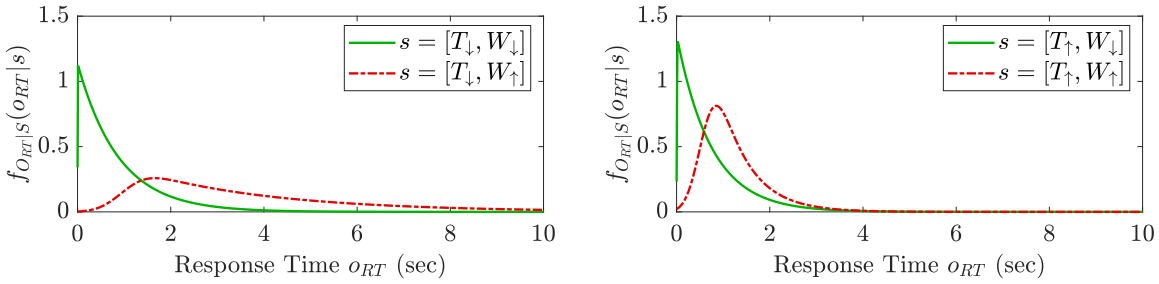


Figure A.14. Emission probability function $\mathcal{E}_W(o_{RT}|s_T, s_W)$ for workload in the coupled-emission model. The left diagram (a) shows the emission probabilities when the trust state is $T_\downarrow$. For Low Trust and Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|T_\downarrow,W_\downarrow}(o_{RT}|T_\downarrow, W_\downarrow)$ is characterized by an ex-Gaussian distribution with $\mu_{T_\downarrow,W_\downarrow} = 0.0018$, $\sigma_{T_\downarrow,W_\downarrow} = 0.0034$, and $\tau_{T_\downarrow,W_\downarrow} = 0.8804$. For Low Trust and High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|T_\downarrow,W_\uparrow}(o_{RT}|T_\downarrow, W_\uparrow)$ is characterized by an ex-Gaussian distribution with $\mu_{T_\downarrow,W_\uparrow} = 0.9845$, $\sigma_{T_\downarrow,W_\uparrow} = 0.4138$, and $\tau_{T_\downarrow,W_\uparrow} = 2.8825$. The right diagram (b) shows the emission probabilities when the trust state is $T_\uparrow$. For High Trust and Low Workload, the response time (o_{RT}) PDF $f_{O_{RT}|T_\uparrow,W_\downarrow}(o_{RT}|T_\uparrow, W_\downarrow)$ is characterized by an ex-Gaussian distribution with $\mu_{T_\uparrow,W_\downarrow} = 0.0063$, $\sigma_{T_\uparrow,W_\downarrow} = 0.0067$, and $\tau_{T_\uparrow,W_\downarrow} = 0.7439$. For High Trust and High Workload, the response time (o_{RT}) PDF $f_{O_{RT}|T_\uparrow,W_\uparrow}(o_{RT}|T_\uparrow, W_\uparrow)$ is characterized by an ex-Gaussian distribution with $\mu_{T_\uparrow,W_\uparrow} = 0.5578$, $\sigma_{T_\uparrow,W_\uparrow} = 0.2603$, and $\tau_{T_\uparrow,W_\uparrow} = 0.6510$.

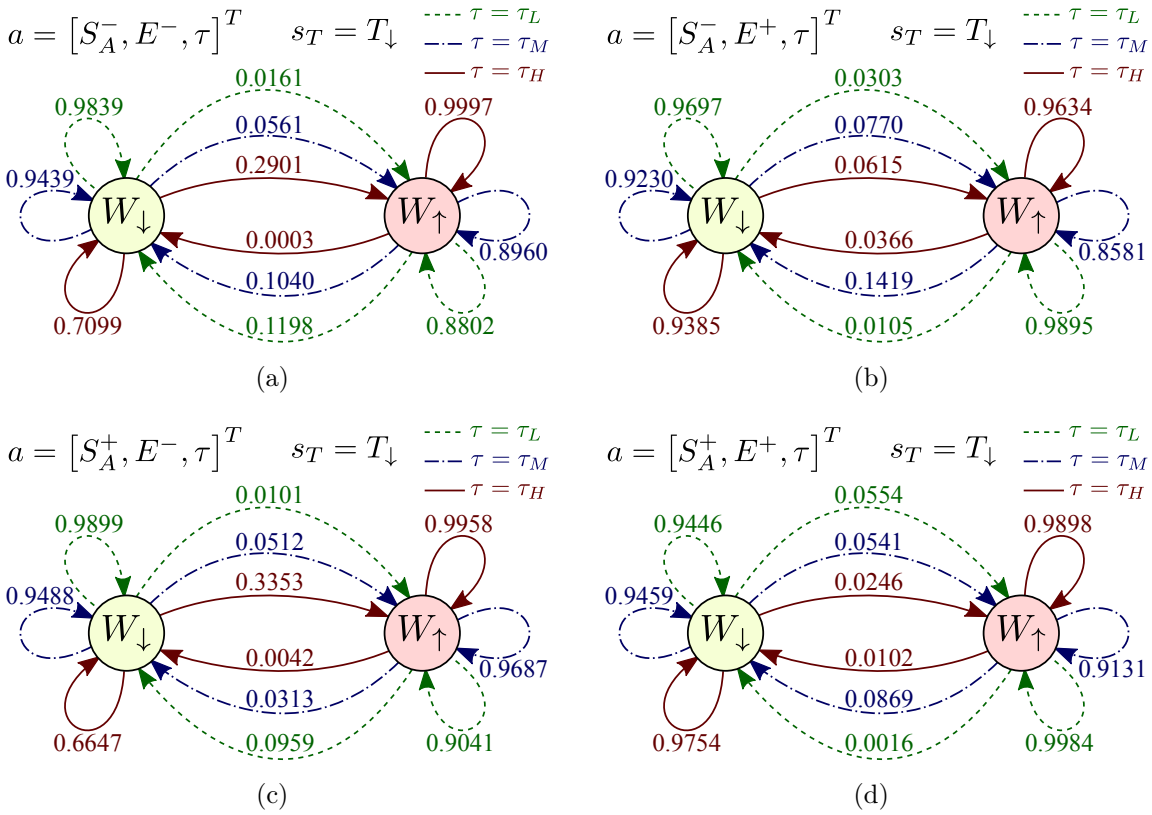


Figure A.15. Transition probability function $\mathcal{T}_W(s'_W|s_T = T_\downarrow, s_W, a)$ for workload in the coupled-emission model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

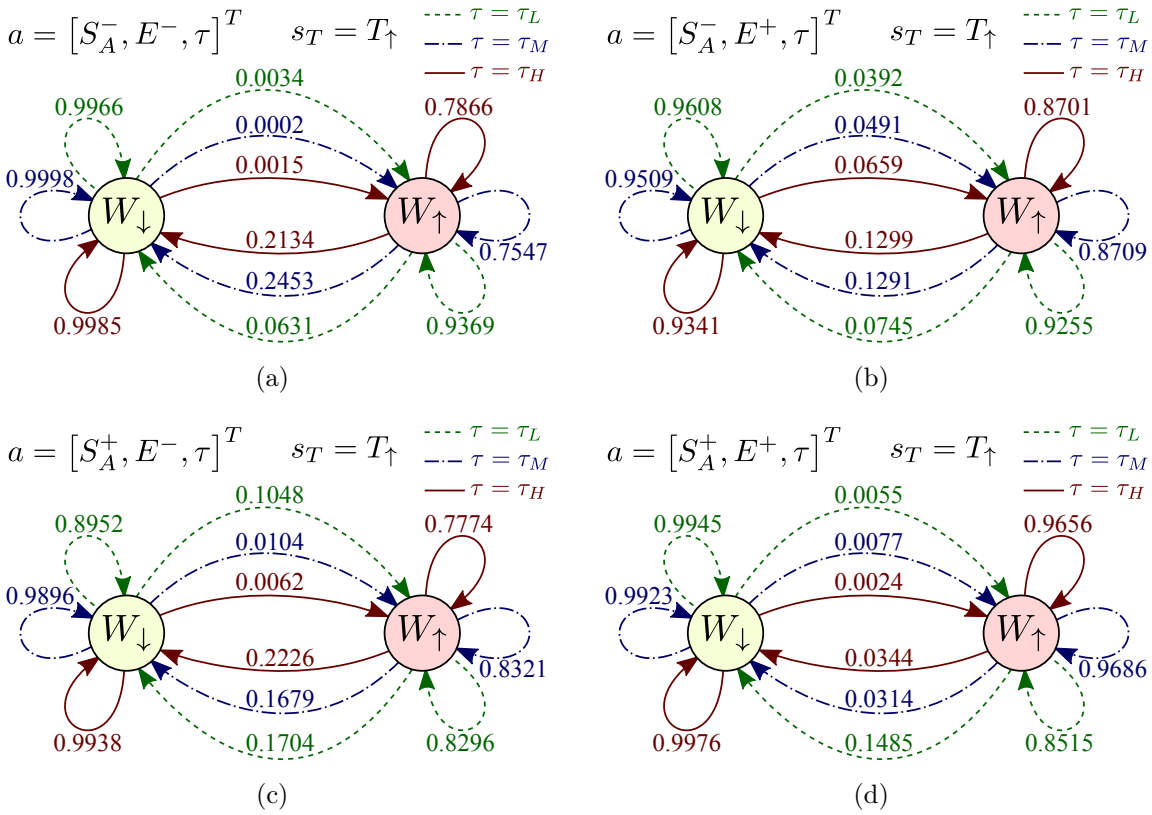


Figure A.16. Transition probability function $\mathcal{T}_W(s'_W|s_T = T_\uparrow, s_W, a)$ for workload in the coupled-emission model. Probabilities of transition are shown beside the arrows. The top-left diagram (a) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Faulty last experience E^-. The top-right diagram (b) shows the transition probabilities when the decision-aid recommends Light Armor S_A^- and the participant had a Reliable last experience E^+. The bottom-left diagram (c) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Faulty last experience E^-. The bottom-right diagram (d) shows the transition probabilities when the decision-aid recommends Heavy Armor S_A^+ and the participant had a Reliable last experience E^+.

B. CONTROL POLICIES TO VARY AUTOMATION TRANSPARENCY

We present the control policies for the independent, coupled-transition, and coupled-emission POMDP models of human trust-workload behavior discussed in Chapter 4. The control policies are calculated for three values of ζ: 0.50, 0.85, and 0.95. Each of the policies describe the optimal choice of transparency (low transparency τ_L, medium transparency τ_M, or high transparency τ_H) given the belief state estimates of trust and workload stated based on the corresponding model, the current recommendation (stimulus absent S_A^- or stimulus present S_A^+), and the experience based on last reliability of the automation (faulty E^- or reliable E^+).

B.1 Independent Model

We calculate the total reward function $\mathcal{R}$ and the corresponding control policy for three values of reward weights $\zeta = 0.50$, $\zeta = 0.85$, and $\zeta = 0.95$. The control policies corresponding to each of the reward weights for the independent model are depicted in Figures B.1, B.2, and B.3.

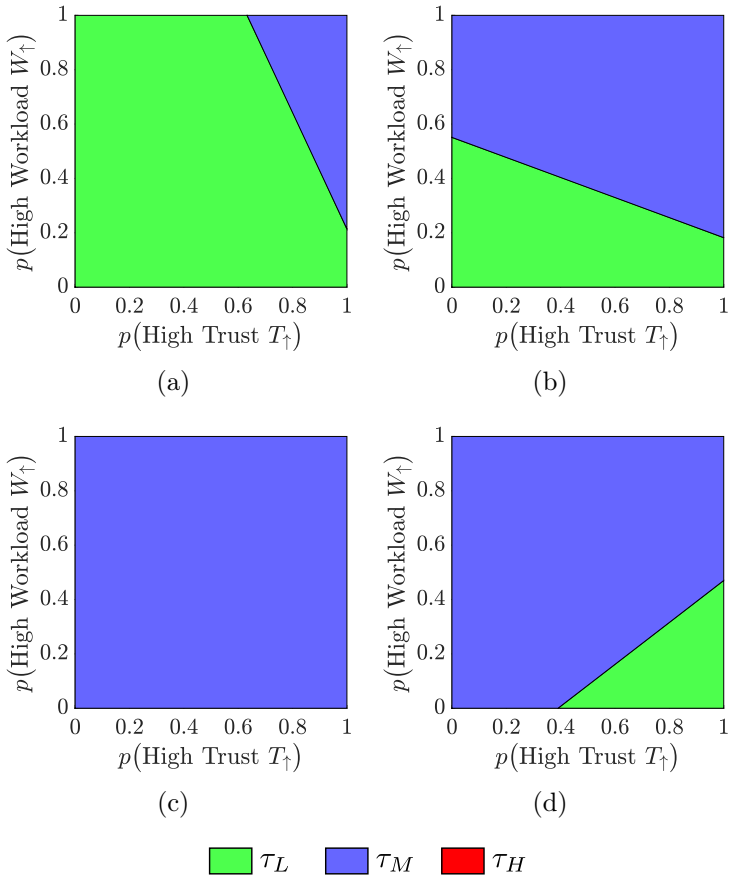


Figure B.1. Closed-loop control policy corresponding to the reward function with $\zeta = 0.50$ for the independent model. In this case, the reward function gives equal importance to the decision and response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

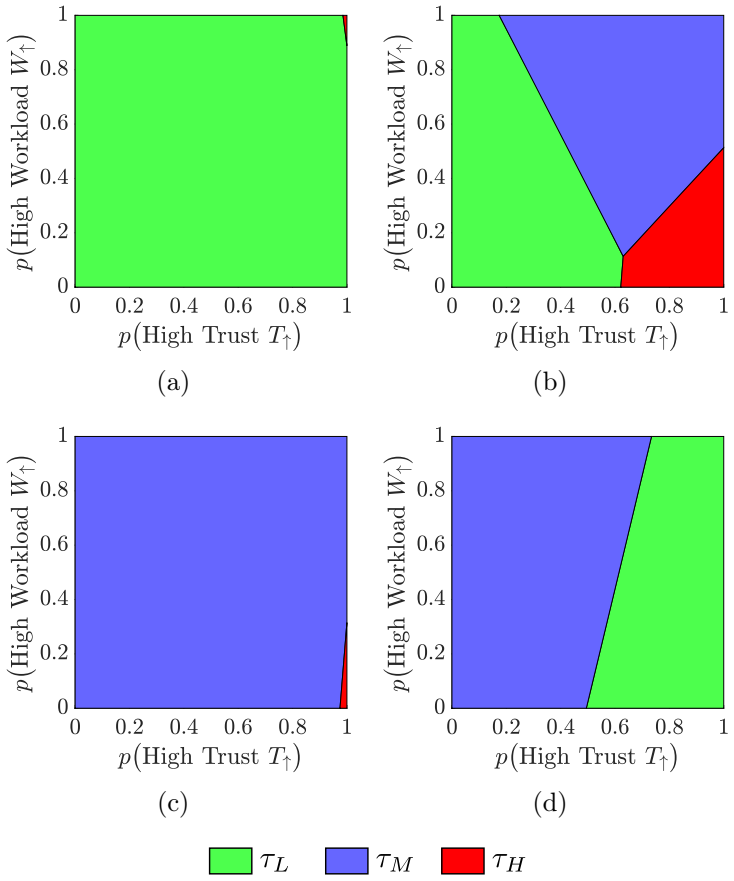


Figure B.2. Closed-loop control policy corresponding to the reward function with $\zeta = 0.85$ for the independent model. In this case, higher importance is given to the decision rewards as compared to the response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

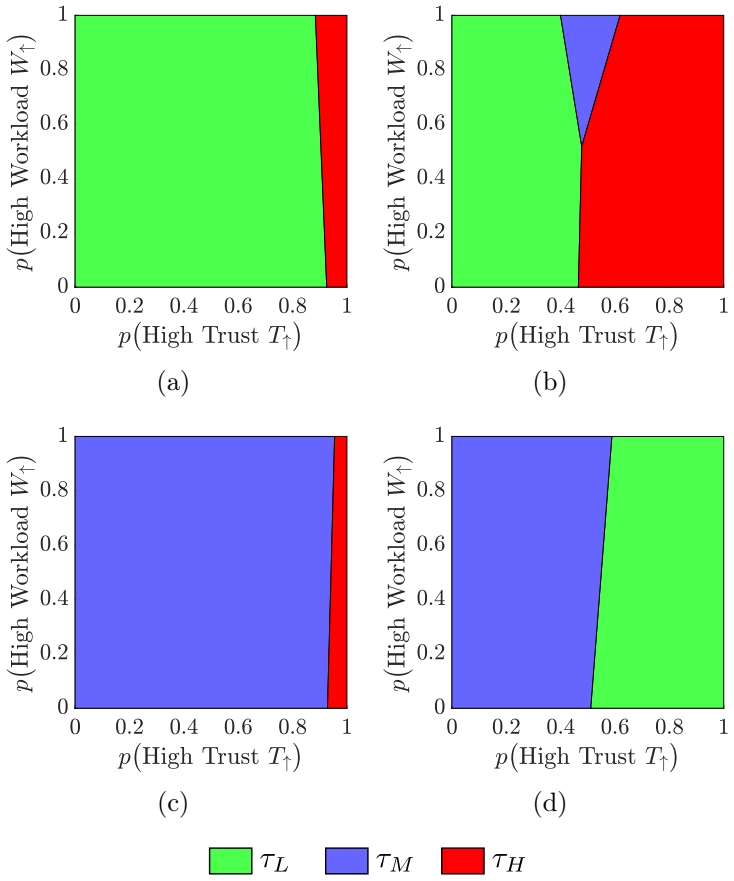


Figure B.3. Closed-loop control policy corresponding to the reward function with $\zeta = 0.95$ for the independent model. In this case, a very high importance is given to the decision rewards as compared to the response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

B.2 Coupled-Transition Model

We calculate the total reward function $\mathcal{R}$ and the corresponding control policy for three values of reward weights $\zeta = 0.50$, $\zeta = 0.85$, and $\zeta = 0.95$. The control policies corresponding to each of the reward weights for the coupled-transition model are depicted in Figures B.4, B.5, and B.6.

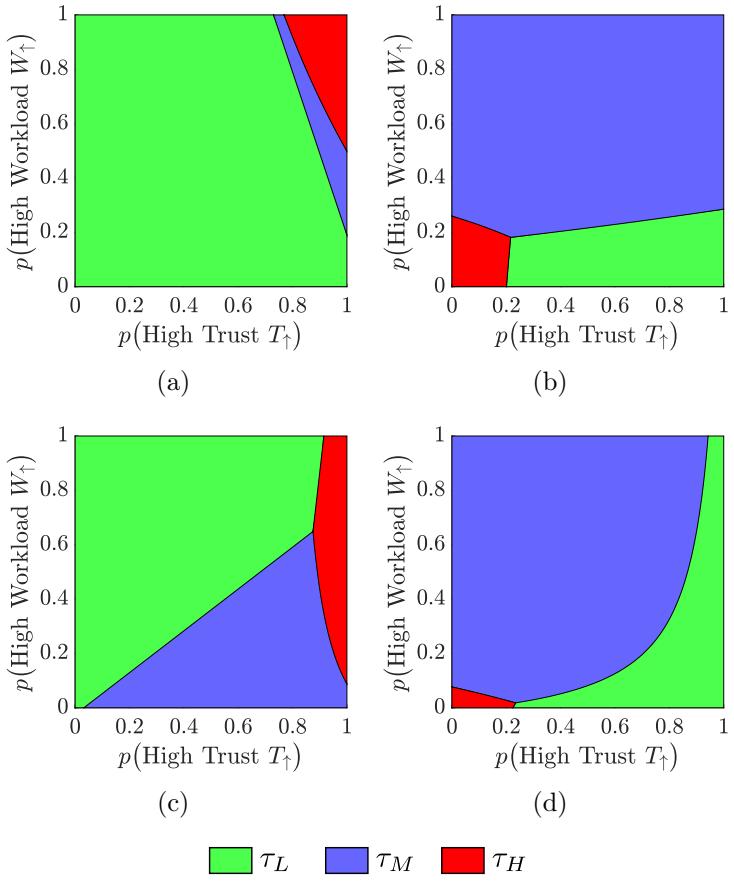


Figure B.4. Closed-loop control policy corresponding to the reward function with $\zeta = 0.50$ for the coupled-transition model. In this case, the reward function gives equal importance to the decision and response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

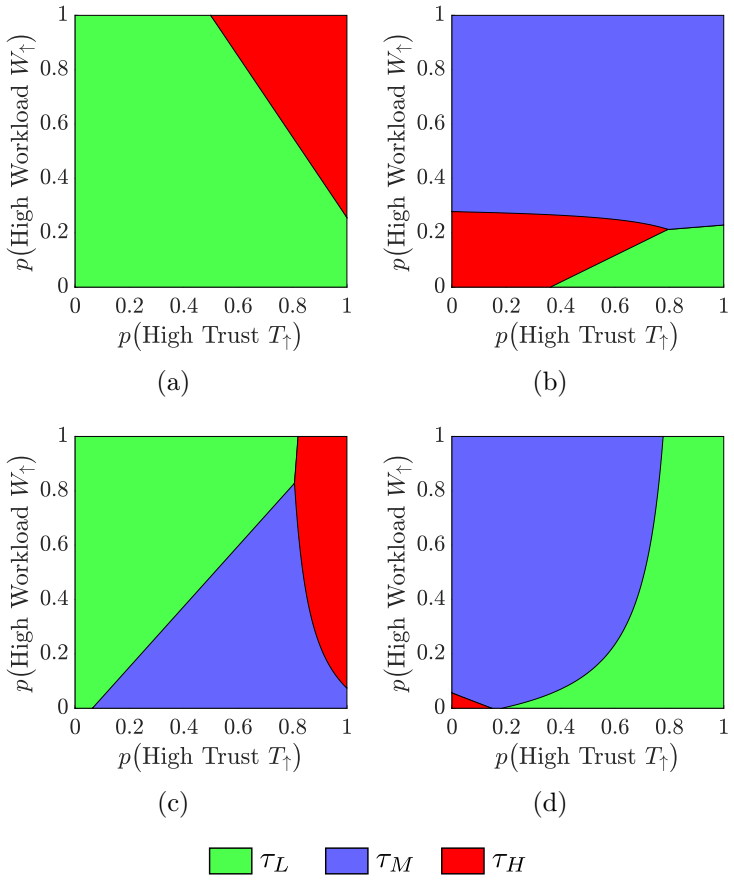


Figure B.5. Closed-loop control policy corresponding to the reward function with $\zeta = 0.85$ for the coupled-transition model. In this case, higher importance is given to the decision rewards as compared to the response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

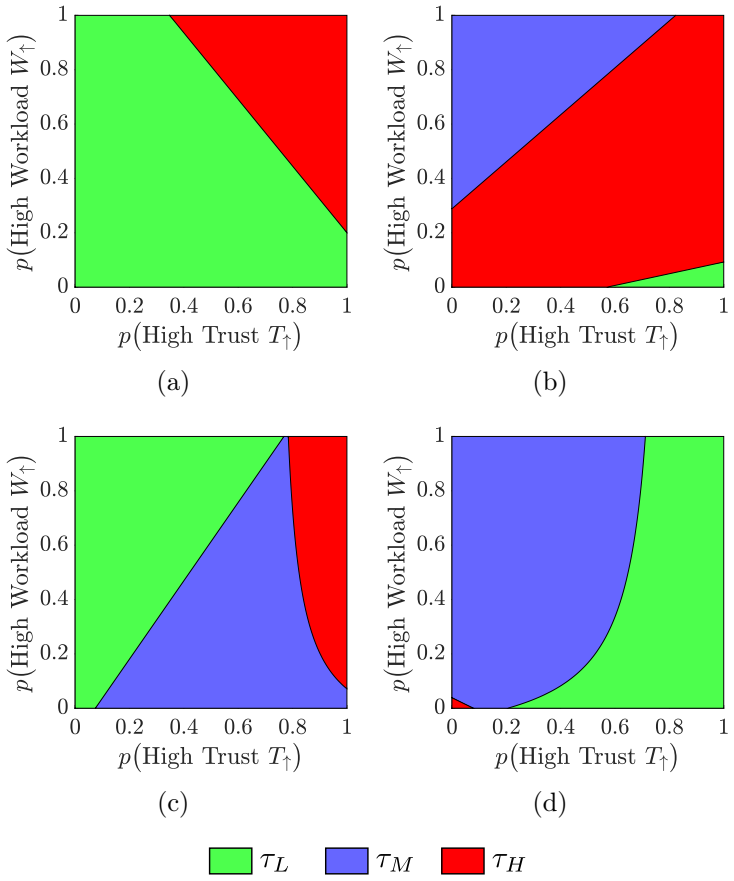


Figure B.6. Closed-loop control policy corresponding to the reward function with $\zeta = 0.95$ for the coupled-transition model. In this case, a very high importance is given to the decision rewards as compared to the response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

B.3 Coupled-Emission Model

We calculate the total reward function $\mathcal{R}$ and the corresponding control policy for three values of reward weights $\zeta = 0.50$, $\zeta = 0.85$, and $\zeta = 0.95$. The control policies corresponding to each of the reward weights for the coupled-emission model are depicted in Figures B.7, B.8, and B.9.

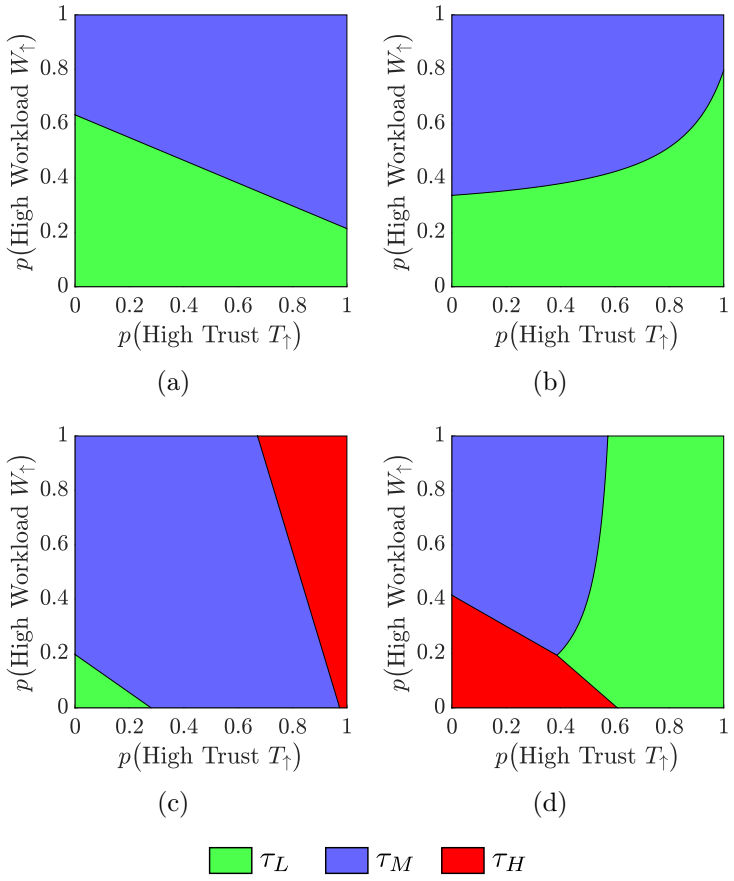


Figure B.7. Closed-loop control policy corresponding to the reward function with $\zeta = 0.50$ for the coupled-emission model. In this case, the reward function gives equal importance to the decision and response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

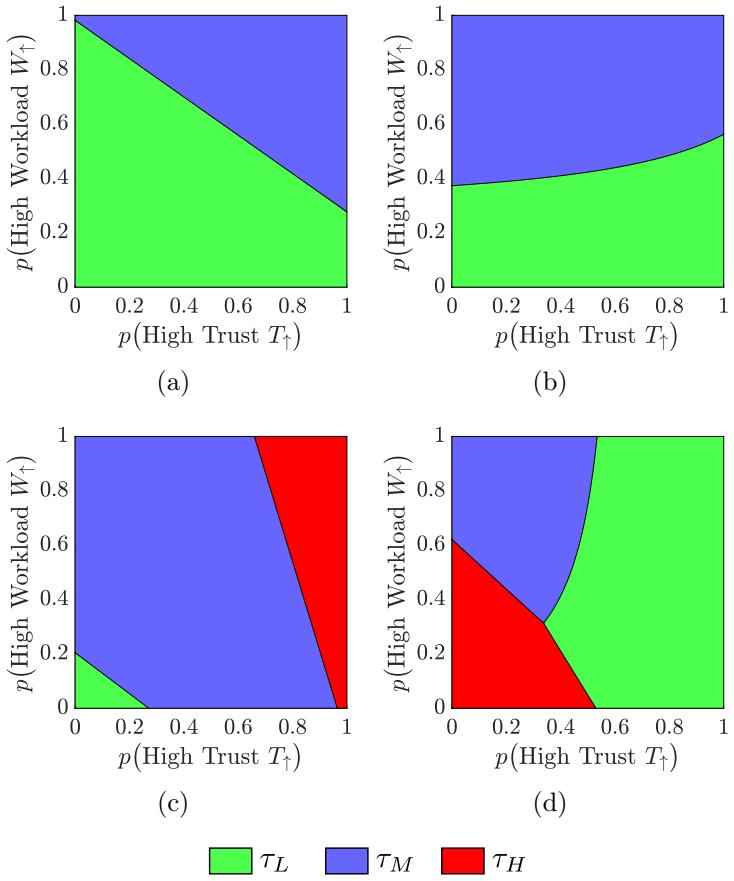


Figure B.8. Closed-loop control policy corresponding to the reward function with $\zeta = 0.85$ for the coupled-emission model. In this case, higher importance is given to the decision rewards as compared to the response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

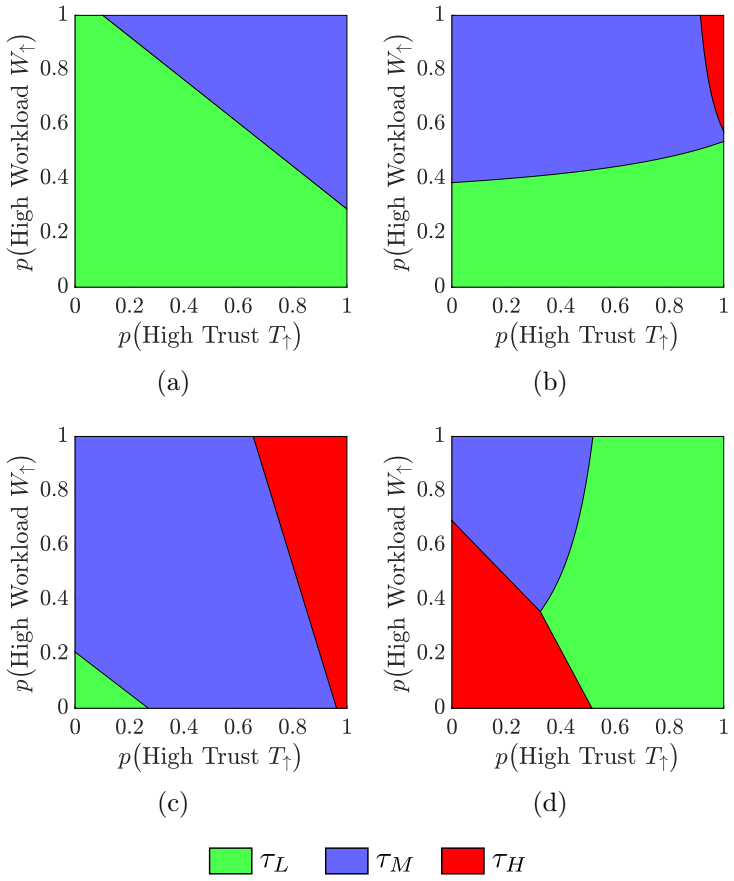


Figure B.9. Closed-loop control policy corresponding to the reward function with $\zeta = 0.95$ for the coupled-emission model. In this case, a very high importance is given to the decision rewards as compared to the response time rewards. Subfigure (a) corresponds to $a_{S_A} = S_A^-, a_E = E^-$, (b) corresponds to $a_{S_A} = S_A^-, a_E = E^+$, (c) corresponds to $a_{S_A} = S_A^+, a_E = E^-$, and (d) corresponds to $a_{S_A} = S_A^+, a_E = E^+$.

VITA

Kumar Akash

EDUCATION

Ph.D. in **Mechanical Engineering** August 2020
Purdue University, West Lafayette, IN, USA
GPA: 4.00/4.00
Advisor: Dr. Neera Jain
Thesis: Reimagining Human-Machine Interactions through Trust-based Feedback

M.S. in **Mechanical Engineering** May 2018
Purdue University, West Lafayette, IN, USA
GPA: 3.93/4.00
Advisor: Dr. Neera Jain

B.Tech. in **Mechanical Engineering** May 2015
Indian Institute of Technology, Delhi, India
GPA: 9.09/10.00
Advisor: Dr. Sudipto Mukherjee
Thesis: Growth Plate Preserving Intramedullary Nail for Pediatric Patients

RESEARCH INTERESTS

Deterministic and Probabilistic Dynamic Modeling; Optimal Control; Human-Machine Interaction; Machine Learning

RELEVANT COURSEWORK

Human Factors in Engineering, Machine Learning, System Identification, Multidisciplinary Design Optimization, Optimal Control & Estimation, Adaptive Control, Nonlinear Feedback Control, Hybrid Systems

AWARDS AND RECOGNITION

- **Bilsland Dissertation Fellowship**, Purdue University, Aug 2019. Awarded by the Dean of the Graduate School to provide support to outstanding Ph.D. candidates in the final year of doctoral degree completion.
- **Batch of Sixty Seven (BOSS) Award**, IIT Delhi, May 2015. Awarded for the best experimental project in mechanical engineering discipline submitted during the session 2014–2015.
- **Summer Undergraduate Research Award**, IIT Delhi, May 2013. Awarded by Industrial Research and Development Unit of IIT Delhi for exceptional research potential displayed at the undergraduate level.

SKILLS

- Software: MATLAB, Simulink, NI LabVIEW (Real-Time & FPGA), SolidWorks, Unreal Engine
- Languages: Python, C++, JavaScript, LaTeX

INDUSTRIAL EXPERIENCE

1. **Research Intern** May 2019–August 2019

   Honda Research Institute, San Jose, CA, USA

   Adaptive Transparency Framework for Level 2 Autonomous Driving

   - Analyzed the effects of augmented reality-based transparency cues on driver's cognitive states using eye-tracking, galvanic skin response, and manual takeover tendencies.
   - Developed models to capture the dynamic effects of transparency on driver's cognitive states.
   - Established optimal control policies to improve driving performance by dynamically varying transparency based on driver's estimated cognitive states.

2. **Mechanical Intern** May 2014–July 2014

   Dover India Innovation Center, Bangalore, India

   Designing a 15000 lbf Hydraulic Planetary Winch for Tulsa Winch Group, Oklahoma, USA

   - Reinforced design decisions using structural analysis of critical components to optimize the weight-to-strength trade-off.
   - Designed hydraulic circuit of the winch along with selection and validation of hydraulic motor required to power the hydraulic winch.

RESEARCH EXPERIENCE

1. **Graduate Research Assistant** August 2015–August 2020

   Purdue University, IN, USA

   Reimagining Human-Machine Interactions through Trust-based Feedback

   - Designed multiple human subject studies to analyze human trust and workload behavior during interactions with an automated decision-aid and collected data using in-person experiments as well as using online experiments conducted through Amazon Mechanical Turk.
   - Created machine-learning-based and control-oriented models to estimate and predict human trust and workload based on human behavior.
   - Developed a classification-based framework to estimate human trust using extracted features from psychophysiological signals including electroencephalogram (EEG) and galvanic skin response (GSR).
   - Synthesized optimal control algorithms that enable machines to respond to changes in human trust in real time to improve human-machine collaboration and validated closed-loop performance through human subject experimentation.

2. **Undergraduate Researcher** July 2014–May 2015

   Indian Institute of Technology Delhi, India

   Growth Plate Preserving Intramedullary Nail for Pediatric Patients

- Modeled and simulated fractured pediatric femur bone.
- Designed a segmented nailing solution with sufficient rigidity to support human body weight.
- Analyzed the stability of the bone and implant system, followed by prototyping and testing for design validation.

3. **Undergraduate Researcher** January 2014–January 2015

   Indian Institute of Technology Delhi, India

   Design of Internal Hub Gear for Bicycles

   - Developed a robust and economical 2-speed hub gear system for Hero Cycles Limited, India's largest cycle manufacturer, to be used in cycles for rural areas.
   - Conceptualized two separate designs based on epicyclic gears; one to be mounted on the rear-wheel and the other on pedals.
   - Validated the designs by simulations followed by prototyping and testing.

4. **Undergraduate Researcher** January 2013–May 2014

   Indian Institute of Technology Delhi, India

   Design and Development of Active Magnetic Bearing System

   - Developed a frictionless and lubricant free bearing and its controller by implementing a magnetically levitated rotor.
   - Optimized the core design of the electromagnet and developed a prototype of the active magnetic bearing system.
   - Designed a Result Adaptive PID controller algorithm that could be heuristically tuned and validated it using a National Instruments Real-Time and FPGA controller.

TEACHING EXPERIENCE

1. **Systems, Measurements, and Control (ME 365)** Fall 2019

   Purdue University, IN, USA

   Substitute lecturer; covered the topic of noise characterization and reduction.

2. **Systems, Measurements, and Control (ME 365)** Fall 2015
   Purdue University, IN, USA
   Graduate teaching assistant for one laboratory section with an enrollment of 22 undergraduate students.

3. **Design of Machines (MCL 211)** Spring 2015
   Indian Institute of Technology Delhi, India
   Undergraduate teaching assistant for the course with an enrollment of 180 undergraduate students.

PROFESSIONAL ACTIVITIES

Journal Reviewer

- IEEE Transactions on Control Systems Technology 2019–Present
- IEEE Transactions on Human-Machine Systems 2017–Present
- IEEE Access 2017–Present

Conference Reviewer

- IEEE Conference on Intelligent Transportation Systems 2020–Present
- IFAC Conference on Cyber-Physical & Human Systems 2018–Present
- American Control Conference (ACC) 2016–Present
- IEEE Conference on Decision and Control (CDC) 2016–Present

Professional Society Memberships

- American Society of Mechanical Engineers (ASME)
- Institute of Electrical and Electronics Engineers (IEEE)

PUBLICATIONS

Journal Articles

- **Kumar Akash**, Tahira Reid, and Neera Jain, "Dynamic Coupling of Human Trust and Workload in Human-Machine Interactions." (In Preparation)
- **Kumar Akash**, Griffon McMahon, Tahira Reid, and Neera Jain, "Human Trust-based Feedback Control: Dynamically varying automation transparency to optimize human-machine interactions." *IEEE Control Systems Magazine*, 2020. (Accepted)
- **Kumar Akash**, Wan-Lin Hu, Neera Jain, Tahira Reid, "A Classification Model for Sensing Human Trust in Machines Using EEG and GSR," *ACM Transactions on Interactive Intelligent Systems*, 2018. doi: 10.1145/3132743
- Wan-Lin Hu, **Kumar Akash**, Tahira Reid, Neera Jain, Tahira Reid, Neera Jain, "Computational Modeling of the Dynamics of Human Trust During Human-Machine Interactions," *IEEE Transactions on Human-Machine Systems*, 2018. doi: 10.1109/THMS.2018.2874188

Conference Articles

- **Kumar Akash**, Neera Jain, and Teruhisa Misu, "Towards Adaptive Trust Calibration for Level 2 Driving Automation," in *22nd ACM International Conference on Multimodal Interaction*, Oct. 2020. (Submitted)
- Nayara Faria, Coleman J Merenda, Richard Greatbatch, Kyle Tanous, Chihiro Suga, **Kumar Akash**, Teruhisa Misu, and Joseph L Gabbard, "The Effect of Augmented Reality Cues on Glance Behavior and Driver-initiated Takeover in Conditionally Automated Driving," in *12th International ACM Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI)*, Sept. 2020. (Submitted)
- **Kumar Akash**, Katelyn Polson, Tahira Reid, and Neera Jain, "Improving Human-Machine Collaboration Through Transparency-based Feedback — Part I: Human Trust and Workload Model," in *2nd IFAC Conference on Cyber-Physical & Human Systems*, Dec. 2018.
- **Kumar Akash**, Tahira Reid, and Neera Jain, "Improving Human-Machine Collaboration Through Transparency-based Feedback — Part II: Control Design

and Synthesis," in *2nd IFAC Conference on Cyber-Physical & Human Systems*, Dec. 2018.

- **Kumar Akash**, Tahira Reid, and Neera Jain. "Adaptive Probabilistic Classification of Dynamic Processes: A Case Study on Human Trust in Automation." In *2018 Annual American Control Conference (ACC)*, pp. 246-251. IEEE, 2018.
- **Kumar Akash**, Wan-Lin Hu, Tahira Reid, and Neera Jain. "Dynamic modeling of trust in human-machine interactions." In *2017 American Control Conference (ACC)*, pp. 1542-1548. IEEE, 2017.
- Wan-Lin Hu, **Kumar Akash**, Neera Jain, and Tahira Reid, "Real-Time Sensing of Trust in Human-Machine Interactions," in *1st IFAC Conference on Cyber-Physical & Human Systems*, pp. 48-53. Dec. 2016.

Invited Talks

1. **Kumar Akash**, Tahira Reid, and Neera Jain, "A Classification Model for Sensing Human Trust in Machines Using EEG and GSR." ACM Intelligent User Interfaces (IUI) Conference 2019, Los Angeles, CA, March 16-20, 2019.
2. **Kumar Akash**, Tahira Reid, and Neera Jain, "Reimagining Human-Machine Interactions Through Trust-Based Feedback." Student Lightning Talks, 2019 Southwest Robotics Symposium, Arizona State University, Tempe, AZ, January 24-25, 2019.