Review: Mathematical Models for Trust in Human-Automation Interactions

LUCERO RODRIGUEZ RODRIGUEZ, CARLOS BUSTAMANTE ORELLANA, YUN KANG, ERIN CHIOU, and LIXIAO HUANG, Arizona State University, USA

The introduction of increasingly autonomous systems operating in human environments may require closer attention to how people trust such systems if the promise of better performance and safety is to be realized. Trust in automation is a rich and complex process that has sparked myriad measures and approaches to study and understand it. Mathematical models have been powerful tools to provide useful insights into dynamical processes of trust in automation. This paper provides a mini-review on how varied mathematical models have been applied and have helped us better understand trust in automation. At last, this paper also suggests dynamical approaches to address multi-time scales of trust in automation dynamics and how machine learning and dynamical modeling should be combined to study the topic.

CCS Concepts: • **Dynamical Models of Trust in Automation** ; • **Modeling Trust in Automation** → *Performance Dynamics*; Reliance Dynamics; Risk Dynamics;

Additional Key Words and Phrases: Trust, Reliance, Decision Making, Risk Dynamics, Dynamical Models

1 INTRODUCTION

Rapid advances in automation technologies, including autonomous vehicles, robotics, autonomous web-based systems, and user experience frameworks and decision aids, are dramatically impacting almost every aspect of our daily life. Understanding how humans work with automation is vital for automation to work most beneficially for humans. *Trust*, which is not always uniformly defined, has been identified as a key factor influencing human-automation interactions [17, 19, 21, 27].

In the literature, there are variations in the definition of trust. Some researchers define trust as an attitude or expectation such as Rotter (1967) [30], who defines interpersonal trust as the *"expectancy held by an individual that the word, promise or written communication of another can be relied upon"*. Barber (1983) [2] defined trust with three expectations: *persistence* of the natural physical order, *technically competent performance*, and *fiduciary responsibility*. Rempel et. al. (1985) [29] defines trust in another person as a dynamical expectation that first falls into the stage of predictability, then dependability, and then faith. Mayer et. al (1995) [24] define trust as the *"willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that party"*. This definition identifies that individuals must willingly put themselves at risk. Similarly, Kramer (1999) [18] defines trust as a behavioral result or state of vulnerability or risk, such as *"a state of perceived vulnerability or risk that is derived from an individual's uncertainty regarding the motives, intentions, and perspective actions of others on whom they depend"*. Falcone et. al. (2001)[6] defines trust as a mental state, a belief of a cognitive agent to achieve a desired goal through another agent. The most widely used definition since its publication is that of Lee and See (2004) [21]. Based on Fishbein and Ajzen's framework and idea [7], Lee and See (2004) define reliance as a human behavior and trust as *"the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"* (p. 54) [21].

When it comes to choosing whether to rely on an autonomous agent, trust (an internally held set of expectations and beliefs) precedes reliance (an externally observable behavior). Reliance (choosing to engage an autonomous agent) is a proxy for trust. The two are not perfectly related, although they do generally correlate positively; this relationship has been demonstrated to increase in strength with the increasing complexity of combat autonomy and novelty of situations (e.g., Sanders et al. 2019 [31]). Like trust, reliance is dynamic and changes over time. Because trust is examined as it

correlates with behaviors such as reliance; the strength of the relationship between an operator's trust ratings and behavior can vary widely.

The mentioned definitions and concepts of trust are important to derive mathematical models to further understand trust in automation dynamics. Since the late nineties, researchers started developing models to have a better understanding of the dynamics of trust and reliance on automation. In this mini-review, we aim to analyze how varied mathematical models have been developed since 1992 by adopting those definitions related to trust in automation. We hope that our effort provides a toolbox to the scientific community on how to model trust in automation with different techniques, their limitations, feasibility, and generalizability for different human-automation interactions.

2 MATHEMATICAL MODELS OF TRUST IN AUTOMATION

In this section we will provide a mini-review of mathematical models for trust in automation in the category of using discrete-time modeling approach including the application of decision field theory, step-function's approach, and probabilistic approach.

2.1 Stochastic Difference Equations

Rempel, Holmes, and Zanna (1985) [29] proposed that trust is a dynamic attitude that follows a particular sequence of dimensions to form gradually over time; and they identified predictability, dependability, and faith as the three dimensions that influence an individual's acceptance of a trustee to form the basis of trust. This concept was subsequently applied to mathematical models of trust in automation by Lee and Moray (1992&1994) [19, 20], Muir (1994&1996)[27, 28], and Lee and See (2004)[21].

2.1.1 Lee & Moray Models (1992,1994). Lee & Moray (1992) [19] used linear regression models to examine the factors affecting trust in automation; and dynamical models to explore how trust changes over time. Their model for trust dynamics was validated using data from a laboratory experiment which consisted in participants operating a simulated orange juice pasteurization plant. In this experiment, each participant reported their trust level (on a scale from 1 to 10) at the end of each of 60 trials recorded over three days. Lee & Moray (1992) [20] started with the development of linear regression models for factors affecting trust. The fitting of these linear regression models suggested that self-reported trust was influenced by two factors, namely the occurrence of fault in a system, and the performance of such system as measured by the total output efficiency (total output/total input). Unfortunately, the linear regression models proposed by Lee & Moray fail to reflect the dynamic response of trust to these variables. The linear regression equation simply predicts trust as a linear combination of the current level of performance and the fault in the system, without regard for the past occurrence of a fault, the past values of performance, or the past values of trust. To give information about the memory of trust, and/or the effect of past occurrences of faults and performance, Lee & Moray (1992&1994) [19, 20] proposed and use a stochastic difference model for trust $T(t)$ which depends on the performance $P(t)$ and fault $F(t)$ of the automation in current trial t and previous trial at $t-1$.

$$T(t) = \phi_1 T(t-1) + \underbrace{A_1 P(t) + A_1\phi_2 P(t-1)}_{\textbf{Performance}} + \underbrace{A_2 F(t) + A_2\phi_3 F(t-1)}_{\textbf{Occurrence of faults}} + a(t) \tag{1}$$

where A_1 is the weighting of system performance, A_2 is the weighting of the occurrence of fault, ϕ_1 and ϕ_2 are time constants of the autoregressive moving average vector (ARMAV) model, and $a(t)$ is a random noise perturbation at trial t. The work of Lee & Moray (1992) provides a first step towards modelling trust between humans and machines, and its influence on operators' control strategies and decision making. Trust model (1) assumed that trust dynamics at trial t is a linear combination of (1) trust T at previous trial $(t-1)$; (2) performance P of the system at trial t and $t-1$; (3)

the occurrence of a fault F at trial t and $t-1$; and (4) environment perturbation $a(t)$ that is out of operator's control. Items (2) and (3) can be linked to the risk assessment of using automation when the participant makes a decision to use automation or not.

One limitation of model (1) is that it fails to explain and test why trust is in fact a linear combination of previous trust, performance, and fault. The applications of trust model (1) may be too restricted, i.e., it can be applied to the specific mentioned experiment as it needs the input of performance and faults. To have a more general model, there is a need to define performance, faults, and other factors such as workload, risk that impact the decision of using automation.

2.1.2 Muir & Moray Models (1994,1996). Muir (1994) [27] adopted the trust definition from Barber (1983) [2] to have the following definition of trust in automation

"*Trust (T) is the expectation (E), held by a member of a system (i), of persistence (P) of the natural (n) and moral social (m) orders, and of technically competent performance (TCP), and of fiduciary responsibility (FR), from a member (j) of the system, and is related to, but is not necessarily isomorphic with, objective measures of these properties.*" which can be summarized as the equation below:

$$T_i = \underbrace{E_i(P(n,m))}_{\textbf{Expectation of persistence}} + \underbrace{E_i(TCP_j)}_{\textbf{Expectation of technically competent performance}} + \underbrace{E_i(FR_j)}_{\textbf{Expectation of fiduciary responsibility}} \tag{2}$$

where T is a composite expectation, comprised of the three expectations: P is the fundamental expectation of persistence; TCP includes skill-, rule-, and knowledge-based behaviours; and FR includes notions of intention, power and authority. The trust model (2) suggests simple linear impacts from persistence (P), technically competent performance (TCP), and fiduciary responsibility (FR), when in fact a multiplicative model or a more complex model may turn out to be a more accurate mathematical representation. For example, the three component expectations need not be equally important; each component expectation may have to be weighted according to its importance in a particular context, and may have nonlinear impacts. Muir proposed the following hypothetical regression model of human trust in a human or machine referent,

$$\begin{aligned} T_i = \quad & B_0 + B_p E_i(P(n,m)) + B_t E_i(TCP_j) + B_f E_i(FR_j) + B_{pt} E_i(P(n,m)) E_i(TCP_j) + \\ & B_{pf} E_i(P(n,m)) E_i(FR_j) + B_{tf} E_i(TCP_j) E_i(FR_j) + B_{ptf} E_i(P(n,m)) E_i(TCP_j) E_i(FR_j) \end{aligned} \tag{3}$$

where B_i are parameters. According to Muir's (1994) model (2), based on Rempel et al.'s (1985) [29] stage model, Muir and Morray in 1996 [28] showed that predictability should be the best predictor of overall trust early in an operator's experience, followed later by dependability and then faith. Thus, they proposed the following model that was applied to their experiment data

$$Trust = Predictability + Dependability + Faith \tag{4}$$

which was extended to the following full model by including three additional components

$$Trust = Predictability + Dependability + Faith + \textbf{Competence} + \textbf{Responsibility} + \textbf{Reliability}. \tag{5}$$

Muir's models (2) and (5) are extended from Barber's [2] and Rempel et al.[29]'s models and showed that the perceived predictability is one of the bases of trust, which in turn, is the foundation for an operator to make an estimate about the future behavior of a referent. The accuracy of that prediction may be assessed by comparing it with the actual behavioral outcome. Besides, a person who makes an estimate may associate a particular level of confidence with such

an estimate. Hence, confidence is a qualifier which is related to a particular estimate; it is not synonymous with trust. One biggest limitation is how to define and measure those three components: Predictability, Dependability, Faith.

2.1.3 Busemeyer & Townsend Model (1993). Decision Field Theory (DFT) is a dynamic-cognitive approach to human decision-making based on psychological rather than economic principles [3]. This type of model has been used to understand the evolution of the preferences among options of a human decision-maker [22]. DFT provides a mathematical approach to understand the cognitive and motivational mechanisms that guides humans in the process of decision making within a changing environment. Busemeyer & Townsend (1993) [4] applied DFT to the decision making of using automation or not. There are two options: rely on automation (A) or use manual control (M). Let S_1 and S_2 correspond to uncertain events, where S_1 is the occurrence of an automation fault and S_2 is the occurrence of a fault that compromises manual control. Variables y_{Mj} and y_{Aj} are the possible payoffs if event S_j occur, where $j = 1, 2$. This model has two basic Subjective Expected Utility (SEU) functions given by,

$$V_A(n) = W(S_1)u(y_{A1}) + W(S_2)u(y_{A2})$$
$$V_M(n) = W(S_1)u(y_{M1}) + W(S_2)u(y_{M2})$$

where $u(y_{Mj})$ and $u(y_{Aj})$ are the utilities of the payoff, $W(S_j)$ is the subjective probability weight (attention given to event S_j) and n is the n^{th} sample. Note that $\sum_j W(S_j) = 1$. Busemeyer & Townsend (1993) [4] define $P(n)$ as the weighted preference state of choosing action automation A over manual M, and they assumed that $P(n)$ is determined by two factors: the previous state of preference $P(n-1)$ and the valence difference of $V_A(n) - V_M(n)$. Thus $P(n)$ is defined as below,

$$P(n) = (1-s)P(n-1) + [V_A(n) - V_M(n)] = (1-s)P(n-1) + d + \epsilon(n) \tag{6}$$

where $V_A(n) - V_M(n) = d + \epsilon(n)$ has an average of d and the related residual $\epsilon(n)$ which represents the change in valence difference produced by the moment-to-moment fluctuations in attention during deliberation. The parameter s is the growth-decay rate, which determines the influence of the previous preference state $P(n-1)$. DFT offers an appropriate modeling approach to describe the decision to adopt automatic or manual control. The preference of automation over manual model (6) has been extended and applied in modeling of trust and self-confidence in Gao-Lee Model 2006 [8] and Maanen and Dongen (2005) [32].

2.1.4 Gao & Lee Model (2006). Gao and Lee (2006) [8] use Lee and See (2004) [21] definition of trust which identifies it as a factor influencing decision making. Gao and Lee first provided the formulation of the extended DFT model (EDFT) on the preference dynamics $P(n)$,

$$P(n) = (1-s)P(n-1) + s \times d + \epsilon(n) \tag{7}$$

which is essentially an autoregressive model that considers a linear combination of the previous preference state $P(n-1)$ and the new input on the current preference state d in an uncertainty environment described by $\epsilon(n)$. The modeling approach of $P(n)$ has been applied to develop a quantitative model of trust and self-confidence that is linked to decision making in automation usage. The trust T and self-confidence SC take the following forms,

$$\begin{aligned} T(n) &= (1-s)T(n-1) + s \times B_{CA}(n) + \epsilon(n) \\ SC(n) &= (1-s)SC(n-1) + s \times B_{CM}(n) + \epsilon(n) \end{aligned} \tag{8}$$

where B_{CA} and B_{CM} are the input for the evolution of trust and self-confidence, representing the fact that automation and manual control capabilities are the primary factors influencing the operator's decision to rely on automation or use manual control. The belief in the automation's capability (B_{CA}) or the operator's manual capability (B_{CM}) are constructed through a piece-wise function that utilize Fishbein et. al. framework where beliefs represent information

base that determines attitudes, and attitudes determine intentions and consequently behaviors [7, 8]. Let $\epsilon_P(n)$ be a random variable with zero mean and variance σ_P^2. The authors define the preference of A over M $P(n)$ as the difference between trust and self-confidence

$$P(n) = T(n) - SC(n) = (1-s)P(n-1) + s \times [B_{CA}(n) - B_{CM}(n)] + \epsilon_P(n) \tag{9}$$

which characterizes multiple sequential decisions instead of the single decisions addressed by DFT. This model is based on psychological principles and depicts the dynamic interaction between the operator and automation. This dynamic interaction describes the relationship between the operator, state of the automation, and the interface where the operators receives information. The model replicates empirical results on inertia of trust and nonlinear relationship between trust, self-confidence, and reliance. The authors acknowledge two limitations to the model. First, it was assumed that the automation and operator capabilities are available, which are the primary input variables. Second, the fit obtained for the model validation is not enough given that the fit has to be done with a greater range of experimental data. A useful feature of the basic DFT, and extended DFT, model is that it does not contain an explicit variable for risk but it is taken into account through the SEU functions. Additionally, the model has the potential to be generalizable to other task environments where there is a human-automation interaction.

2.1.5 van Maanen et. al. Model (2005). Maanen and Dongen (2005) [32] used Falcone et. al. (2001) [6] definition of trust and referred it as a mental state and belief of a cognitive agent i about the achievement of a desired goal through another agent j or through agent i itself. Maanen and Dongen (2005) [32] implemented the framework of Decision Field Theory (DFT) to derive a model for task allocation where both the human and machine act together as a team. The model contains four mathematical definitions of: task execution state, trust state, allocation preference state, and preferred task execution state. The task execution and preferred task execution state are sequences of characters, while trust state and allocation preference state are real numbers ($\in \mathbb{R}$). The authors consider that trust depends on past experiences. Thus, they use the agent's task execution state to update the trust state. Furthermore, the model assumes that preferences are determined by trust in the self and trust in the other. Then, the allocation preference state is given by the difference of trust states corresponding to self-trust and trust in the machine. Finally, the previous states helps the operator make a preferred decision on the allocation task, which updates the preferred task execution state. The authors propose an experimental design to validate the model, where the goal is to predict, as a human-machine team, the location of a disturbance which can only occur at one of three locations. The authors do not clarify what would be the payoff of a correct or incorrect task allocation, which are essential for the SEU functions in the DFT model. Hence, the assessment of risk in the model is not clearly defined. This model may not be generalizable for more complex environments where the operator faces more than three different environmental factors that can influence several different outcomes.

2.1.6 Akash et. al. Model (2017). Akash et. al. (2017) [1] proposed a three-state model for trust in automation (T) by adopting the modeling work of Jonker and Treur (1999) [16] and the concept of trust of Hoff and Bashir (2015) [10] who classified the trust into the following three categories:

1. Dispositional trust is based on characteristics of the human such as culture, gender, age, and personality.
2. Situational trust consists of those factors that are external to the human (e.g., task difficulty) and those that are internal to the human (e.g., domain knowledge).
3. Learned trust is based on the accumulation of experiences with autonomous systems and influences the initial mindset of the human.

The trust model of Jonker and Treur (1999) [16] described the change in trust to be proportional to the difference of experience and trust. Akash et. al. (2017) [1] adapted Jonker's model and introduced two additional states—Cumulative Trust (C_T) and Expectation Bias (B_X)—to accommodate the bias in human behavior due to human's perception of past trust and their expectations as follows:

$$\begin{aligned} T(n+1) - T(n) &= \alpha_e \left[E(n) - T(n)\right] + \alpha_c \left[C_T(n) - T(n)\right] + \alpha_b \left[B_X(n) - T(n)\right] \\ C_T(n+1) &= \left[1-\gamma\right] C_T(n) + \gamma T(n) \\ B_X(n+1) &= B_X(n) \end{aligned} \tag{10}$$

where α_e, α_c, and α_b are called the experience rate factor, cumulative rate factor, and bias rate factor, respectively. Additionally, γ discounts older trust levels faster, and thus it can be called the trust discounting factor. The specific assumptions of modeling trust in the model (10) are that the change in trust $T(n+1) - T(n)$ linearly depends on three terms: (a) the difference between experience and present trust $E(n) - T(n)$, (b) the difference between cumulative trust and present trust $C_T(n) - T(n)$, and (c) the difference between expectation bias and present trust $B_X(n) - T(n)$. If the present experience is less than the present trust level, then the predicted trust level decreases and vice-versa. The cumulative trust was defined as an exponentially weighted moving average of past trust level, so that it includes the learned trust in the model using a weighted history of past trust levels. The expectation bias, which accounts for a human's expectation of a particular interaction with an autonomous system, is intended to be constant during an interaction but it can change between different interactions. Akash et. al. (2017) model (10) is difficult to generalize for other scenarios because it is based on specifically asking participants whether they trust the automation or no. It assumes there is a way of measuring current levels of trust in order to predict the future trust, so the prediction of trust based on real-time behaviors is not possible with this model.

2.2 Step Functions Approaches

Itoh and Tanaka in 2000 [13] define trust as the expectation or belief that the automation is dependable. Their definition of trust evolves from Rempel's definition [29], who says that trust evolves from predictability, dependability and faith. This modeling basis is in line with the definition used in Muir and Moray Models (1994&1996) [27, 28]. Itoh and Tanaka [13] proposed a step functional model for trust in automation by adopting Rempel's definition [29]. The authors define X as the universal set of all possible automation's operating conditions, where $x, y \in X$ and $x < y$ means that y is more difficult than x. The model of trust is composed of three different sets, the faithful condition (F), the dependable condition (D), and the predictable condition (P). Additionally, the sets UF, UD and UP refer to the complement of the faithful (F), dependable (D) and predictable condition (P), respectively. The Itoh-Tanaka model in 2000 [13] defines trust as a function of the automation's operating condition (x). Theire model takes the following form,

$$t(x) = \begin{cases} 1 & \text{if} \quad x \in D \\ 0 & \text{if} \quad x \in UD \quad \text{or} \quad x \in UF \\ \alpha_x & \text{if} \quad x \in UP \end{cases} \tag{11}$$

where $\alpha_x \in [0, 1]$ and is dependable of the operator's personality and/or their experience with the automation. Several experimental studies on trust in automation measure trust via survey given to the participant. This is 10 point subjective scale where a value 1 represents *not trusting the automation at all* and value 10 represents *complete trust on the automation.* This rating is one final measure of trust that is only able to inform us about the overall trust the operator has on the automation, but is unable to give us a major scope of its dynamics. The authors relate their model to this subjective

rating, where the subjective rating of trust is given by,

$$t_s = \frac{\int_X t(x)dx}{|X|} \tag{12}$$

which estimates the subjected trust value after the operator interacts with an automation. This is calculated by taking the mean of the function $t(x)$. This model of trust does not assess a task environment or risk as factors that influences the operator. Additionally, this model is highly general in the description of the dynamics of trust not only for automation but for human's trust in other being or operational device. Additionally, the authors utilized variables such as, faith and operator's personality, but did not described or suggested on methodologies to measure these variables. Given that the authors decided to describe their model as a function of sets, then this model is not generalizable for its use with other added variables such as, risk, environmental factors, and workload from the task given to the operator.

2.2.1 Monir at.al. Model (2020). Monir et. al. [25] define trust using a step function, where the level of trust that a human has on a robot depends on both the human and robot performances. These two types of performances determine which region of trust build-up process the human is in. Such regions were obtained from previous works [13, 26–29] suggesting that trust is commonly captured based on three aspects: predictability, dependability, and faith. Monir et. al. [25] model presented by the following form,

$$T(t) = \begin{cases} 0 & ;\text{for } R_p(t) < f_P \\ \epsilon & ;\text{for } f_P \leq R_p(t) < f_D \\ min(1, \epsilon + tanh(c\Delta P)) & ;\text{for } f_D \leq R_p(t) < f_F \\ 1 & ;\text{for } R_p(t) \geq f_F \end{cases} \tag{13}$$

where T is the level of trust that the human has on the robot, t is the time, R_P is the robot performance, $\Delta P = R_P(t) - f_D$, $f_P = \sigma H_P(t)$, $f_D = \rho H_P(t)$, f_F is the robot performance at which trust reaches its maximum value, H_P is the human performance, σ and ρ are small numbers, and c and ϵ are variables which value depends on the human preferences. In this model, the robot is in the unpredictable region when $R_P(t) < f_P$ as its performance is much lower than the human performance, it's in the predictable region when $f_P \leq R_P(t) < f_D$, it's in the dependable region after $R_P(t) \geq f_D$ where the trust build up process starts, and finally, it's in the faithful region when the trust value reaches it's maximum at $R_p(t) = f_F$.

This model was proposed for the specific scenario of human-robot interaction where a human supervises the correct execution of a classification task made by a robot. The experiment was done in a physical laboratory environment and the task consisted in separating cubes of different colors by placing them into different counters. This model can be applied to other scenarios by making some changes to their definitions and measures. The separation of trust into different regions based on the performance of human and robot made by the authors of this model is applicable to any other scenario of human-robot interaction, however, the way of measuring the robot and human performance may have to change depending on the particular case of analysis.

2.3 Probability Approaches - Bayesian Network

2.3.1 van Maanen et. al Model (2007). Maanen et al (2007)[33] used Lee & See [21] definition of trust, where trust is described as the attitude that an agent will help achieve an individual's goal in a situation characterized by uncertainty and vulnerability; and it is also a convert or cognitive state [6]. Maanen et al (2007) [33] model is based on the idea that the performance of humans in cooperation with aids (agents) and vice versa, perform better than humans and aids working separately. The authors' objective is to verify the possibility for the aid to make more accurate trust

assessments and accurate reliance decisions. Thus, the decision aid calculates its trust in the operator and on itself based on the reliance decision making capabilities each time there is feedback information. In the design of the decision aid, the authors derived a probabilistic model for trust using binomial likelihood function, the Beta probability density function (pdf) and Bayes' rule to update estimations. Since the operator's and aid's behavior can be a 'success' or 'failure', then this can be described by the Bernoulli distribution. Since the Beta distribution is the conjugate prior for the Bernoulli distribution in Bayesian inference, then this saves numerical computation for the posterior in Bayesian inference. Hence, the aid uses Bayes' rule to update its estimations over the different values that the agents' capabilities can have. The aid needs to estimate the probability of a successful outcome in each trial, θ_a^x where $x \in$ {prediction, reliance} and $a \in$ {operator, aid}.

To capture the trust dynamics, participants were asked to perform a pattern recognition task with advice of a decision aid. Taking this into account, the factors that considered in the model is the number of successes and failures for each agent, which refer to the number of correct and incorrect pattern recognition that each agent made. Additionally, the task environment and model does not assess a risk variable. Even though the model could be modified so the human agent does the trust assessment, the model is not generalizable for other human-automation interaction with more complex environments. This model idea could potentially be incorporated as part of another model that enable us to predict trust and reliance in automation.

2.3.2 Xu & Dudek Model (2015)-Dynamic Bayesian Network. Xu and Dunek (2015) [35] use their own definition of *trust*. They defined it as - one's belief in the competence and reliability of another, and defined that the level of reliance induces the level of trust. Xu and Dunek (2015) designed, developed and evaluated a Dynamic Bayesian Network model for a human's level of trust in a robot teammate with the primary assumption being that the operator occasionally intervenes to aid the robot at the given task. Hence, the human operator acts as a "supervisor" of the robot completing a series of tasks. The probabilistic model by Xu and Dunek (2015) [35], Online Probabilistic Trust Inference Model (OPTIMo), formulates Bayesian beliefs over the human's moment-to-moment trust states. The Bayesian model incorporates variable-rate sources of information in a probabilistic manner, and can accommodate an arbitrary belief for previous trust. The model relates the human's latent trust state (t_k) to the robot's task performance ($p_k \in [0, 1]$). It also uses human interventions ($i_k \in \{0, 1\}$), trust change reports ($c_k \in \{-1, 0, +1, \emptyset\}$), absolute trust feedback ($f_k \in \{[0, 1], \emptyset\}$), and e_k the presence of a task change for a time window k.

The model was validated using data from a laboratory simulated experiment; and the authors were able to accurately predict human's trust-induced behaviors. Neither the model or the experiment used for model validation takes into account the risk as a factor. This model differs from others since it uses the operator's behavior of reliance on the automation to imply the level of the operator's trust. This framework has the potential to be generalizable to model trust dynamics for other human-automation interaction test-bed that contains the operator's reliance behavior throughout the trial. Additionally, other variables and type of automation could be added to the model.

2.3.3 Wang et. al. Model (2018). Wang et. al. (2018) [34] use the definition of trust given by Lee & and See [21] to explore if the learning and inference process, of human operators updating their trust based on system reliability over time, approximately follow the principles of Bayesian probabilistic inference. Wang et. al. (2018) [34] started by applying Bayesian probabilistic inference to approximate human operators' perceived automation reliability. They proposed that the system reliability, denoted as r^*, can be estimated by the posterior mean of a Beta distribution with parameters $N_T + M_T + 1$ and $N_F + M_F + 1$, as follows, $r^* = \frac{N_T+M_T+1}{N_T+N_F+M_T+M_F+2}$ where N_T and N_F are the number of successes and failures of the automation, respectively. In addition, M_T and M_F are parameters to model mathematically how strong a prior belief in automation success or failure is, respectively. Subsequently, the authors compared the automation

reliability approximates against values reported by the operators and found high correlations between the two. Finally, moderate correlations were found between the Bayesian estimates and the human operators' reported trust. These findings indicate that human operators' learning and inference process of automation reliability can be approximated by Bayesian inference. This model is difficult to generalize to include other factors that may determine trust in automation and other scenarios. As it is based on the solely probabilistic inference of human operators' perceived automation reliability, including other factors require a complete reformulation of the model.

3 DISCUSSION AND FUTURE DIRECTION

Researchers have made significant contributions to the understanding of factors influencing trust in automation. Lee and See's review (2004) [21] emphasizes how the increasing complexity of automated systems has produced the necessity of understanding how humans trust these systems. Thus, trust cannot be viewed as a static phenomenon or outcome. Instead, trust is an interactive and dynamic process with multiple sources of influence [10, 21]. Hoff and Bashir (2006) [10] reviewed the empirical work that followed Lee and See's [21] (2004) and defined three sources of variability in trust in automation: dispositional, situational, and learned. Dispositional trust is a person's attitude towards autonomous agents generally, based on pre-existing knowledge and demographic characteristics. Learned trust is based on prior experience with a specific autonomous system. Through this experience, operators learn about the system's capabilities, observe its performance, and develop expectations about the system's reliability. Learned trust is impacted by performance; for example, operators judge automation that fails on easy tasks more harshly (greater trust degradation) than agents that fail on difficult tasks [23]. Situational trust takes into account conditions in the environment. This may include the external context, for example, the impact of weather on task difficulty; it may also include internal variables, such as the operator's current mental load or emotional state. Situational variables may change quickly and be difficult to observe, adding complexity to any effort to assess trust. Both learned trust and situational trust can change over short periods of time. Learned trust can change as an operator gains new experience with an agent and it may shift abruptly if an agent suddenly demonstrates a change in behavior or experiences a failure. Situational trust is even more unstable; anything altering the internal or external environment may impact the operator's moment-to-moment trust in an agent – even factors that are not directly related to the agent or its performance. This three-layer trust model is compatible with Mayer et al.'s (1995) [24] human-human model, which considers the trustor's characteristics (dispositional), perceived risk (situational), and perceived trustworthiness that is dynamically updated by observing trustee behavior (learned). Additionally, this is in line with Rempel's trust definition [29], which first falls into the stage of predictability (situational), then dependability stage (learned), and then faith (dispositional). Note that Rempel's definition [29] of trust has been implemented to Lee and Moray (1992 & 1994) [19, 20], Muir (1994 & 1996)[27, 28], and Lee and See (2004)[21].

All mathematical models reviewed in this paper agree that *Trust in Automation* (TiA) is shaped by complicated dynamical processes involving the human operator, automated system, and environment, all of which intertwine and impact each other over time. Current models of trust do not sufficiently specify how the trust measurement is related to the overall model or its components. The disconnect between measurements and models prevents our systematic study of TiA. Thus, there is a need to develop a unified model that can link directly to these measurements. In order to do so, from the mathematical point of view, we could classify the dynamical processes of *Trust in Automation* (TiA) into three timescales:

1 Short-timescale (e.g., in seconds/mins): decision making that is related to predictability and situational trust such as risk,

2 Intermediate-timescale (e.g., in hours/days): reliance and performance dynamics that could be related to dependability and learned trust, and
3 Long-timescale (e.g., in years/decades) that are linked to dispositional trust such as culture.

Lee & Moray's (1992,1994) model can be classified into the intermediate-timescale since it considers the performance and occurrence of fault of the automation in trials recorded over a span of three days. Muir & Moray's (1994, 1996) models would fall into the long-timescale because they consider either the persistence of natural laws or faith which are concepts that involve human previous experiences (which can go back years). Akash et. al (2017) model can be classified into the long-timescale as it considers humans' past experiences (through demographics).Monir et. al (2017) model would fall into the long-timescale too since it considers faith in their approach, however, the model was not validated with data from the participants' past experiences. Thus, this model is classified into the intermediate-timescale. Wang et. al (2018) model uses parameters to model the strength of prior beliefs in the automation success or failure, so it can be classified into intermediate or long-timescales. However, the authors used a dataset where participants' trials were separated by seconds or minutes, so it is classified into the short-timescale. Itoh & Tanaka (2000) model can be considered intermediate and long-timescale since it considers faith (dispositional trust), dependable and predictable conditions which are based on the automation's performance. Models constructed through Decision Field Theory, such as Busemeyer & Townsend (1993), Gao & Lee (2006), and van Maanen et. al. (2005) can be classified on the three timescales since the models are linked to decision making, reliance on automation, performance, and belief (long-timescale). Xu & Dudek (2015) model is a short and intermediate-timescale model since takes into account the automation's performance and the human's decision making in order to calculate what the authors refer to as a trust state.

Now, the natural question would be how to model dynamical processes in these three-time scales? Current research [17] has found that automation performance, workload, and risk are factors that largely affect the decision-making of automation usage, which influences trust and reliance on automation. So, how do we model the decision-making process that occurs on a short timescale? We need to determine what is the risk, workload, and unforeseeable consequences due to stochasticity in the environment. As an example, we can define risk of automation usage at time t being $R(t) = \frac{A}{A+B}$, where A represents task violations (failure) and B represents successful task completions (good performance) with engaged automation over time $[0, t]$.

The recent work by Kohn et. al. (2021) [17] provides a comprehensive narrative review, which addresses known methods that have been used to capture TiA. Trust can be measured by surveys, binary behavioral indicators, sensory-based psychological values, and communication patterns [11, 12]. Surveys measure a person's perception of attitude towards another entity using Likert scales and ordinal data (e.g., Jian, Bisantz, & Drury, 2000 [15]). Often time, survey measures are a one-time event, with the exception that some studies measure it multiple times during a task [5]. Binary behavioral indicators of trust include behavioral indicators: such as use or not use of the automation and for how long, and eye-tracking fixation duration [9, 14]. Sensor-based psychological measures of trust are more nuanced than others, including electroencephalography (EEG), Galvanic skin response (GSR). A useful mathematical model should incorporate both behavioral and physiological data. Due to the complex nature of those measurements and large data sets, there is a need of implementing both machine learning and dynamical modeling approaches to develop powerful mathematical models. Our ongoing research project consists in establishing the foundation of these powerful mathematical models. These models will permit rigorous mathematical analyses and validation with a wide range of human-automation interaction settings.

ACKNOWLEDGMENTS

This research was sponsored by the Army Research Office through Cooperative Agreement Number W911NF-18-2-0271. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. This research was additionally supported through a Research Assistantship awarded by the School of Human Evolution and Social Change from Arizona State University.

REFERENCES

[1] Kumar Akash, Wan Lin Hu, Tahira Reid, and Neera Jain. 2017. Dynamic modeling of trust in human-machine interactions. *Proceedings of the American Control Conference* (2017), 1542–1548. https://doi.org/10.23919/ACC.2017.7963172

[2] Bernard. Barber. 1983. The logic and limits of trust. (1983), 189.

[3] Jerome Busemeyer and Adele Diederich. 2002. Survey of decision field theory. *Mathematical Social Sciences* 43 (07 2002), 345–370. https://doi.org/10.1016/S0165-4896(02)00016-1

[4] Jerome Busemeyer and James Townsend. 1993. Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. *Psychological review* 100 (08 1993), 432–59. https://doi.org/10.1037//0033-295X.100.3.432

[5] Mary Cummings, Lixiao Huang, and Masahiro Ono. 2021. *Investigating the influence of autonomy controllability and observability on performance, trust, and risk perception.* 429–448. https://doi.org/10.1016/B978-0-12-819472-0.00018-6

[6] Rino Falcone and Cristiano Castelfranchi. 2001. *Social Trust: A Cognitive Approach.* Springer Netherlands, Dordrecht, 55–90. https://doi.org/10.1007/978-94-017-3614-5_3

[7] M. Fishbein and Icek Ajzen. 1975. *Belief, attitude, intention and behaviour: An introduction to theory and research.* Vol. 27.

[8] Ji Gao and J.D. Lee. 2006. Extending the decision field theory to model operators' reliance on automation in supervisory control situations. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 36, 5 (2006), 943–959. https://doi.org/10.1109/TSMCA.2005.855783

[9] Gregory Gremillion, Jason Metcalfe, Amar Marathe, Victor Paul, James Christensen, Kim Drnec, Benjamin Haynes, and Corey Atwater. 2016. Analysis of trust in autonomy for convoy operations. 98361Z. https://doi.org/10.1117/12.2224009

[10] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors* 57, 3 (2015), 407–434.

[11] Lixiao Huang, Nancy Cooke, Craig Johnson, Glenn Lematta, Shawaiz Bhatti, Michael Barnes, and Eric Holder. 2020. *Human-Autonomy Teaming: Interaction Metrics and Models for Next Generation Combat Vehicle Concepts.* Technical Report. ARIZONA STATE UNIV EAST MESA AZ MESA.

[12] Lixiao Huang, Nancy J. Cooke, Robert S. Gutzwiller, Spring Berman, Erin K. Chiou, Mustafa Demir, and Wenlong Zhang. 2021. Chapter 13 - Distributed dynamic team trust in human, artificial intelligence, and robot teaming. In *Trust in Human-Robot Interaction*, Chang S. Nam and Joseph B. Lyons (Eds.). Academic Press, 301–319. https://doi.org/10.1016/B978-0-12-819472-0.00013-7

[13] Makoto Itoh and Kenji Tanaka. 2000. Mathematical Modeling of Trust in Automation: Trust, Distrust, and Mistrust. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44, 1 (2000), 9–12. https://doi.org/10.1177/154193120004400103 arXiv:https://doi.org/10.1177/154193120004400103

[14] Q. Jenkins and X. Jiang. 2010. Measuring trust and application of eye tracking in human robotic interaction. *IIE Annual Conference and Expo 2010 Proceedings* (01 2010).

[15] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04 arXiv:https://doi.org/10.1207/S15327566IJCE0401$_0$4

[16] Catholijn Jonker and Jan Treur. 1999. Formal Analysis of Models for the Dynamics of Trust Based on Experiences, Vol. 1647. https://doi.org/10.1007/3-540-48437-X_18

[17] Spencer C. Kohn, Ewart J. de Visser, Eva Wiese, Yi Ching Lee, and Tyler H. Shaw. 2021. Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology* 12, October (2021). https://doi.org/10.3389/fpsyg.2021.604977

[18] Roderick M. Kramer. 1999. Trust and distrust in organizations: emerging perspectives, enduring questions. *Annual review of psychology* 50 (1999), 569–98.

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

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

[21] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 arXiv:https://doi.org/10.1518/hfes.46.1.50$_3$0392 PMID: 15151155.

[22] Seungho Lee, Young-Jun Son, and Judy Jin. 2008. Decision field theory extensions for behavior modeling in dynamic environment using Bayesian belief network. *Information Sciences* 178, 10 (2008), 2297–2314. https://doi.org/10.1016/j.ins.2008.01.009

[23] Poornima Madhavan, Douglas A. Wiegmann, and Frank C. Lacson. 2006. Automation Failures on Tasks Easily Performed by Operators Undermine Trust in Automated Aids. *Human Factors* 48, 2 (2006), 241–256. https://doi.org/10.1518/001872006777724408 arXiv:https://doi.org/10.1518/001872006777724408 PMID: 16884046.

[24] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An Integrative Model of Organizational Trust. *The Academy of Management Review* 20, 3 (1995), 709–734. http://www.jstor.org/stable/258792

[25] Md Khurram Monir Rabby, Mubbashar Altaf Khan, Ali Karimoddini, and Steven Xiaochun Jiang. 2020. Modeling of Trust Within a Human-Robot Collaboration Framework. In *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, Vol. 2020-Octob. https://doi.org/10.1109/SMC42975.2020.9283228

[26] Bonnie Marlene. Muir. 1989. Operators trust in and percentage of time spent using the automatic controllers in a supervisory process control task. (1989).

[27] Bonnie M. Muir. 1994. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* 37, 11 (1994), 1905–1922. https://doi.org/10.1080/00140139408964957

[28] Bonnie M. Muir and Neville Moray. 1996. Trust in automation. part ii. experimental studies of trust and human intervention in a process control simulation. *Ergonomics* 39, 3 (1996). https://doi.org/10.1080/00140139608964474

[29] John K. Rempel, John G. Holmes, and Mark P. Zanna. 1985. Trust in Close Relationships. *Journal of Personality and Social Psychology* 49, 1 (jul 1985), 95–112. https://doi.org/10.1037/0022-3514.49.1.95

[30] Julian B. Rotter. 1967. A new scale for the measurement of interpersonal trust. *Journal of personality* 35 4 (1967), 651–65.

[31] Tracy Sanders, Alexandra Kaplan, Ryan Koch, Michael Schwartz, and P. A. Hancock. 2019. The Relationship Between Trust and Use Choice in Human-Robot Interaction. *Human Factors* 61, 4 (2019), 614–626. https://doi.org/10.1177/0018720818816838 arXiv:https://doi.org/10.1177/0018720818816838 PMID: 30601683.

[32] P.P. van Maanen and K. van Dongen. 2005. Towards Task Allocation Decision Support by means of Cognitive Modeling of Trust. In , *Proceedings of the Eighth International Workshop on Trust in Agent Societies*, C. Castelfranchi, S. Barber, J. Sabater, and M. Singh (Eds.). 168–77. Trust05.

[33] Peter-Paul van Maanen, Tomas Klos, and Kees van Dongen. 2007. Aiding Human Reliance Decision Making Using Computational Models of Trust. In *2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops*. 372–376. https://doi.org/10.1109/WI-IATW.2007.108

[34] Chenlan Wang, Chongjie Zhang, and X. Jessie Yang. 2018. Automation reliability and trust: A Bayesian inference approach. *Proceedings of the Human Factors and Ergonomics Society* 1 (2018), 202–206. https://doi.org/10.1177/1541931218621048

[35] Anqi Xu and Gregory Dudek. 2015. OPTIMo: Online Probabilistic Trust Inference Model for Asymmetric Human-Robot Collaborations. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 221–228.