Faculty of Technology, Natural Sciences and Maritime Studies
Department of Maritime Operations

University of South-Eastern Norway

Modeling Perspective on Human-Automation Interaction (HAI):

Levels and Trust in Automation

Mehdi Poornikoo

A Ph.D. dissertation for the degree of Nautical operations, March 2024

Modeling Perspective on Human-Automation Interaction (HAI): Levels and Trust in Automation

Mehdi Poornikoo

UiT The Arctic University of Norway
Faculty of Science and Technology
Department of Technology and Safety

Norwegian University of Science and Technology
Faculty of Engineering
Department of Ocean Operations and Civil Engineering

University of South-Eastern Norway
Faculty of Technology, Natural Sciences and Maritime Studies
Department of Maritime Operations

Western Norway University of Applied Sciences
Faculty of Business Administration and Social Sciences
Department of Maritime Studies

Doctoral dissertations at the University of South-Eastern Norway no. 204

ISSN: 2535-5244 (print)
ISSN: 2535-5252 (online)
ISBN: 978-82-7206-884-3 (print)
ISBN: 978-82-7206-885-0 (online)

Dedication

In loving memory of my mother, Masoume.

Summary

The advent of Maritime Autonomous Surface Ships (MASS) represents a significant leap forward in the maritime industry, promising to redefine sea transportation's efficiency, safety, and economics. However, this technological advance brings forward the complex interplay between human operators and autonomous systems, particularly in the context of Shore Control Centers (SCCs), where remote operators play critical roles. The success of integrating MASS into the global shipping infrastructure depends not just on technological advancements but equally on understanding and optimizing Human-Automation Interaction (HAI). The transition to supervisory control roles introduces a paradigm shift in operational dynamics. Remote operators are tasked with maintaining oversight over multiple vessels simultaneously, each possibly facing different sea conditions and operational challenges. This multi-vessel management can significantly amplify the cognitive load, requiring operators to prioritize information effectively and make swift decisions to ensure safety and efficiency. One of the primary concerns is the risk of over-reliance on automation, which may lead to complacency and reduced situational awareness. The remote nature of operation may exacerbate these issues, as operators are removed from the immediate physical environment of the vessels they control. Moreover, the unpredictable and dynamic nature of maritime environments makes complete autonomy a challenging goal; remote operators must be prepared to take control in complex or emergency situations.

To address these challenges and leverage the full potential of MASS, it is imperative to develop scientific and robust models of HAI. These models should account for the unique demands of maritime environments and the specific roles of remote operators. By understanding the cognitive, psychological, and social factors that influence remote operators' performance, researchers and practitioners can design more intuitive and effective interfaces and decision-support systems. Effective HAI models can guide the development of training programs tailored to the needs of remote operators, focusing on critical skills such as situational awareness, decision-making under uncertainty, and effective communication with autonomous systems. Moreover, these models can help identify potential sources of error, the operators' responses, and cognitive overload, enabling the design of systems that support operators' decision-making processes and reduce the likelihood of accidents. Two pivotal aspects of these models are the Levels of Automation (LOAs) and Trust in Automation (TiA). Understanding and accurately modeling these dimensions are crucial for designing systems that effectively balance human supervisory control of autonomous capabilities.

In response to the growing scrutiny regarding the validity of Human Factors and Ergonomics (HFE) models, as well as the need for flexible yet credible HAI models, this dissertation concentrated on the importance of models and modeling within Human-Automation Interaction (HAI), particularly emphasizing Trust in Automation (TiA) and Levels of Automation (LOA) as central themes for modeling exploration. This dissertation commences by exploring the significance of scientific modeling and developing criteria that can be utilized to assess the relative scientific credibility of various models. Furthermore, models of Trust in Automation (TiA) were assessed against these criteria not only to showcase the use of the criteria but also to understand the TiA modeling efforts in the literature. On the other hand, epistemological

accounts of modeling efforts were investigated, to realize the suitability of each approach for modeling HAI. The findings suggested simulation as a viable approach to tackle the complexities in modeling TiA and LOA within the context of HAI and supervisory control of MASS. By incorporating models of Trust in Automation (TiA) and Levels of Automation (LOA), simulation offers a powerful tool for examining complex interactions and dynamics that are difficult, if not impossible, to study in real-world settings due to safety, cost, and practicality concerns.

Acknowledgment

As I present this work, I am compelled to express my gratitude to those who have been instrumental in my PhD journey. Undertaking this doctoral work would not have been possible without the guidance and support of my supervisors: Professor Kjell Ivar Øvergård and Associate Professor Frøy Birte Bjørneseth. Thank you both for your assistance throughout this process. Kjell, your mentorship has been invaluable, and your friendship, accompanied by enlightening discussions on "good" science over whiskey and wine, has been a source of great support and inspiration.

My heartfelt thanks go to my friends, whose kindness and companionship made this journey endurable. To Sanda and Kristoffer, for welcoming me(us) from the very beginning, sharing their table, and treating me(us) like family members. To Easa, my FocusMate, FIFA partner, and dear friend. To Ana, Lena, and Boban, for friendship and companionship—Hvala!

I would also like to extend my appreciation to the colleagues at IMA who impacted the PhD environment for the better. To Steven, Munim, Karina, Carina, Halvor, Monica, Fred, Kenn, Tor Erik, Anne H., Tor Inge, Christian H., Paul Nikolai, Morten, Per H., Erlend, Erik Andre, Lene and Per Eirik; thank you!

To my fellow PhD candidates, Karen, Laura, Mariia, Hasan, Amit, Koen, Simen, and William, your camaraderie and engaging discussions were invaluable. A special thanks to my friend and office mate, Mari, for your unwavering support and courageous spirit.

I am profoundly thankful to my father, my solid, steady, foundational rock. Thank you for your unconditional love and support. I aspire to become like you.

Lastly, my deepest gratitude goes to my partner in crime, my beloved wife, Veronica. Your constant encouragement and support have been crucial in this journey. Thank you for always being by my side.

March 2024
Mehdi

List of Publications

Appended Articles

Article 1

Poornikoo, M., & Øvergård, K. I. (2023). Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation. Theoretical Issues in Ergonomics Science, 1-37. https://doi.org/10.1080/1463922X.2023.2233591

Article 2

Poornikoo M., Mansouri M. (2023), "Systems approach to modeling controversy in Human factors and ergonomics (HFE)," 18th Annual System of Systems Engineering Conference (SoSe), Lille, France, 2023, pp. 1-8,
https://doi.org/10.1109/SoSE59841.2023.10178634

Article 3

Poornikoo, M., & Øvergård, K. I. (2022). Levels of automation in maritime autonomous surface ships (MASS): A fuzzy logic approach. Maritime Economics & Logistics, 24(2), 278-301. https://doi.org/10.1057/s41278-022-00215-z

Article 4

Poornikoo M., Gyldensten W., Vesin B., Øvergård, K. I. (In review) Trust in Automation (TiA): simulation model, and empirical findings in supervisory control of Maritime Autonomous Surface Ships (MASS), International Journal of Human-Computer Interaction

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
AIC	Akaike Information Criterion
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
AOI	Area of Interest
ATMS	Air Traffic Management System
AUTC	Area Under Trust Curve
BWM	Best Worst Method
CLD	Causal Loop Diagram
CSE	Cognitive System Engineering
DP	Dynamic Positioning
ECDIS	Electronic Chart and Display and Information System
EDFT	Extended Decision Field Theory
FIS	Fuzzy Inference System
FL	Fuzzy Logic
GNSS	Global Navigation Satellite System
HAI	Human-Automation Interaction
HFE	Human Factors and Ergonomics
HITL	Human-In-The-Loop
HOTL	Human-On-The-Loop
HPM	Human Performance Modeling
IMO	International Maritime Organization
JSC	Joint Cognitive System
MASS	Maritime Autonomous Surface Ship
MCDM	Multi-criteria Decision Making
MUNIN	Maritime Unmanned Navigation through Intelligence in Networks
NDS	Nonlinear Dynamic System
OPTIMO	Online Probabilistic Trust Inference Model
PC	Personal Computer
RQ	Research Question
SCC	Shore Control Center
SD	System Dynamics
SEM	Subject Matter Expert
TCAS	Traffic Collision Avoidance System
TiA	Trust in Automation
TRA	Theory of Reasoned Action

Table of Contents

1 Introduction

1.1 Establishing the Context

Industry 4.0 marks a paradigm shift from conventional automation technologies, which acted as supplements to human work, to a new era of machine autonomy. In this age, automation emerges as an intelligent entity, capable of executing complex tasks such as planning and decision-making, powered by advancements in the Artificial Intelligence (AI) (Aiello et al., 2020; Sepehri et al., 2022). Yet, the vision of machines functioning autonomously within unpredictable and unstructured environments is far from realization. Beyond the technical challenges, there exist credible arguments, such as ethical concerns and the need for accountability, that might limit the absolute autonomy of machines (Coito, 2021; Jordan, 2019). Contemporary research has thus pivoted towards hybrid interaction frameworks that harness both human expertise and automated efficiency. One such framework is supervisory control, where the level of autonomy afforded to machines becomes higher, but human control and supervision are also essential. In such a scenario, the human operator transitions to a supervisory role, monitoring automated functions and intervening as necessary, particularly in unforeseen circumstances or to revise objectives. One of the key applied domains experiencing this transition is the Maritime Autonomous Surface Ship (MASS). Various research and industry projects have anticipated that autonomous ships will become a reality in seas in the coming years (Jalonen et al., 2017; Jokioinen et al., 2016; Laurinen, 2016). The shift towards autonomous shipping is expected to occur gradually, progressing from lower levels of autonomy to higher ones (Laurinen, 2016; Utne et al., 2017), where unmanned ships would likely operate with constrained autonomy, either supervised or controlled by a Shore Control Center (SCC) operator (Porathe et al., 2014; Ringbom et al., 2017.; Rodseth et al., 2018; Rodseth, 2017; Rødseth et al., 2021). This implies that MASS does not equate to completely unmonitored operations, and humans are still required to supervise and analyze the operations performed by autonomous systems.

As automation continues to reshape the working environments, the role of the operator's cognitive factors such as mental workload, trust, and self-confidence, during interactions with automated systems becomes crucial for effective human-automation interaction (Hussein et al., 2020; Lee & See, 2004; Peters et al., 2015). These cognitive factors significantly influence an individual's willingness to use and rely on automation (Gao & Lee, 2006; Hancock et al., 2013; Lee & Moray, 1994; Riley, 1996). Consequently, creating automated systems that respond to the user's cognitive state could enhance task performance and learning (Hancock et al., 2013). In this line, cognitive models may become important in explaining and predicting cognitive states and would enable systems to modify their responses, adjusting transparency, behavior, and autonomy levels (Alonso & De La Puente, 2018; Chen et al., 2015, 2015). Modeling Human-Automation Interaction (HAI) can help design more intuitive, efficient, and safe automated systems by creating a deeper understanding of how humans interact with such systems. This endeavor not only enhances system performance but also ensures that technological advancements align with human well-being and operational safety. A well-constructed model provides a basis for understanding and predicting the complex behaviors of

the operator engaging with the automation. Through modeling, researchers and designers can explore various scenarios, identify potential pitfalls, and develop strategies to enhance interaction and collaboration between human operators and automated systems (Hancock et al., 2013). This approach can significantly reduce the likelihood of unintended consequences, enabling designers to refine automation features based on predictive analyses of human behavior. In essence, cognitive theories and models serve as fundamental tools in bridging the gap between theoretical knowledge and practical application, ensuring that automated systems are not only technologically advanced but also human-centric.

Over the past three decades of HAI research, several models and frameworks have been developed to elucidate the cognitive factors and their effects on human performance and decision-making in interaction with automation (Boubin et al., 2017; Hoff & Bashir, 2015; Parasuraman & Riley, 1997). These models and frameworks often emphasize the importance of trust and levels of automation within the HAI context as the determinants of automation use and reliance (Endsley, 2018; Muir & Moray, 1996; Parasuraman et al., 2000). Despite the prevalent theories and cognitive models, critiques (e.g., Dekker & Hollnagel, 2004; Flach, 1995) have raised significant concerns regarding the scientific credibility and practical utility of the existing frameworks. These critical perspectives form the foundational impetus for this doctoral dissertation. This dissertation is designed to delve into the heart of these issues from multiple vantage points, aiming to uncover and address the underlying challenges. The forthcoming sections will offer an expanded insight into the core problem, delineate the specific objectives of this dissertation, and articulate the research questions that will guide this scholarly investigation.

1.2 Problem Description

The rapid advancement of technologies and their swift adoption as 'work facilitators' by industry stakeholders have triggered continuous changes in the workplace. The essence of human work has undergone significant transformations since the early 20th century. This evolution is evident when comparing perspectives on human-work studies from Scientific Management (Taylor, 1911), Human Factors Engineering (Fitts, 1951), Ergonomics and Cognitive Ergonomics (Meister, 2000; Meister & Enderwick, 2001), Cognitive Systems Engineering (Hollnagel & Woods, 1983), and Human-Computer Interaction (Card et al., 1983). Unlike fields such as mathematics and the natural sciences, where theoretical issues often drive progress, the challenges in Human Factors and Ergonomics (HFE) arise from its practical aspects. The field faces the difficulty of a continuously moving target – practical needs that grow so quickly that they are hard to accurately identify and address.

Despite widespread consensus on the significance of theory in Human-Automation Interaction (HAI) and Human Factors and Ergonomics (HFE) in general, some scholars have noted a concerning shortfall of theoretical grounding in HFE and HAI research (Salas, 2008). Chung (2017) discussed that although HFE is a science-based discipline and its efficacy hinges on solid scientific underpinnings, much of the scientific basis of HFE remains underdeveloped. Salas (2008) further pointed out that, despite having well-established theories in areas such as human information processing, decision-making, and team effectiveness, the field of HFE is

still predominantly atheoretical. The neglect of theoretical frameworks and models in HAI research has been highlighted as a serious issue by various scholars. Salas (2008) warned that an overly applied focus on engineering and design in HFE has led to theories being "*ignored, misused, or abused*." (p.352). Similarly, Hockey (2008) argued that overlooking theoretical foundations could undermine the effectiveness of practical applications in the field. De Greene (1980) noted significant conceptual challenges in managing ergonomics research, particularly emphasizing the problems arising from the reliance on pure static models. This concern is shared by other scholars (e.g., Guastello, 2017, 2023; Karwowski, 2012; Thatcher et al., 2020), who also expressed concerns about the use of static models in understanding cognition and human-automation interactions in dynamic settings. Woods and Dekker (2000) articulated these concerns more intensely. They pointed out that the rapid pace of technological change and the growing scope of technological advancement have made traditional models and methods increasingly inadequate. These models and methods are viewed as oversimplifications that can hinder understanding and progress. According to Woods and Dekker (2000), the reliance on outdated and oversimplified models and methods could ultimately undermine the credibility of the ergonomics field.

Generally, scientific models in human-technology system studies fall into two categories: componential and systemic, each addressing different aspects of human-technology interactions (Øvergård, 2008). Componential models assume system behavior is predictable from its components' behavior, leading to an understanding of the whole system as a simple addition of its parts' behaviors (Card et al., 2005; Dekker, 2005; Hollnagel & Woods, 1983). This approach, rooted in the information-processing view of human cognition, separates mind and body and treats the environment as a passive element (Fodor, 1983; Gardner, 1985; Ihde, 2002). However, the growing complexity of contemporary technical systems has challenged the adequacy of this view (Dekker & Hollnagel, 2004; Perrow, 1999; Woods & Dekker, 2000). Conversely, the systemic view, or Cognitive Systems Engineering (CSE) (Hollnagel & Woods, 1983, 2005), emerged in response to the componential perspective. It views sociotechnical systems as comprising interconnected humans, technology, and the environment, functioning in a goal-directed manner (Vicente, 1999). This approach emphasizes the functional unity of the system, where outcomes are emergent from coordinated activities across components, rather than traits of individual parts (Hollnagel, 2003). Under this perspective, the human operator is an integral part of the Joint Cognitive System (JCS), adapting and responding contextually within the system (Hollnagel & Woods, 1983, 2005).

These viewpoints indicate a growing interest within the HFE community in the need for more robust, adaptable models and methodologies that can keep pace with the evolving landscape of technology and human interaction. The critique suggests a push towards more novel, advanced approaches that can better capture the complexities inherent in modern HFE research and design. However, a foundational step requires a thorough comprehension of the discipline's unique challenges, requirements, and the current models' effectiveness in addressing these needs. Establishing this foundational understanding is crucial before determining which types of models are most appropriate for addressing HFE challenges and outlining strategies for their development.

1.3 Research Focus, Objectives, and Questions

This dissertation investigates the intricate facets of Human-Automation Interaction (HAI), with a particular focus on the Levels of Automation (LOAs) and Trust in Automation (TiA). LOAs and TiA are central in assessing the effectiveness and adoption of automation technologies. The primary objective of this doctoral dissertation is to explore the essential modeling criteria for HAI research within socio-technical frameworks, to clarify the epistemological underpinnings of various modeling methodologies, and to evaluate the adequacy of current modeling efforts in meeting these criteria. Moreover, this dissertation suggests the value of simulation-based modeling techniques as promising solutions capable of overcoming the shortcomings in existing models.

Figure 1 presents an overall research approach of this dissertation, integrating concepts from Human Factors and Ergonomics (HFE), particularly focusing on Cognitive Ergonomics and its application to Human-Automation Interaction (HAI). This approach is designed to evaluate and model trust in automation (TiA) and levels of automation (LOA) within the context of Human-Automation Interaction (HAI) and supervisory control. Key elements discussed in this dissertation are outlined as follows:

1. *Cognitive Ergonomics Domain:* At the core, this dissertation explores cognitive ergonomics with an emphasis on the psychological aspects of HAI. It considers environmental, automation-specific, and individual factors that influence Trust in Automation (TiA). The dynamics of TiA are captured through a feedback loop that includes the individual's expectation of outcomes and their perception of actual outcomes.
2. *Human-Automation Interaction (HAI):* The dissertation investigates the interaction between humans and automation and the crucial role of the levels of automation (LOA) which dictates how humans supervise and control automated systems. As depicted, Trust in Automation (TiA) directly influences the interactions with automated systems and also is influenced by the outcome of the interactions.
3. *Characteristics of HFE & Phenomenon:* The dissertation outlines the broad characteristics of HFE as a discipline, including its scientific, applied, systems, and complex nature. Additionally, it addresses the systems as being linear vs. nonlinear, and open vs. closed, highlighting the diverse contexts in which HFE research is conducted.
4. *HFE Model Evaluation Criteria:* To assess the efficacy of models within HAI, the dissertation suggests a set of scientific evaluation criteria including testability/falsifiability, predictive power, explanatory power, empirical adequacy, pragmatic adequacy, humans as active agents, and dynamic properties of models. These criteria are essential for ensuring that models are both scientifically rigorous and practically relevant.
5. *Modeling Epistemological Assumptions:* The dissertation recognizes the significance of underlying epistemological assumptions in modeling efforts. It differentiates between variance (static) and process (dynamic) models, acknowledging that each has its strengths and applications within HFE. This distinction is crucial for selecting the appropriate modeling methodology based on the nature of the phenomena being studied.
6. *Simulation Modeling:* As a solution to the limitations identified in traditional HFE models, the dissertation proposes simulation modeling as a versatile tool for understanding complex

HAI phenomena. Specifically, it suggests employing Fuzzy Logic Inference Systems and System Dynamics modeling to capture the dynamics of levels and trust in automation. These simulation methodologies offer the flexibility and depth needed to model cognitive states and their evolution over time, providing a more holistic understanding of HAI.

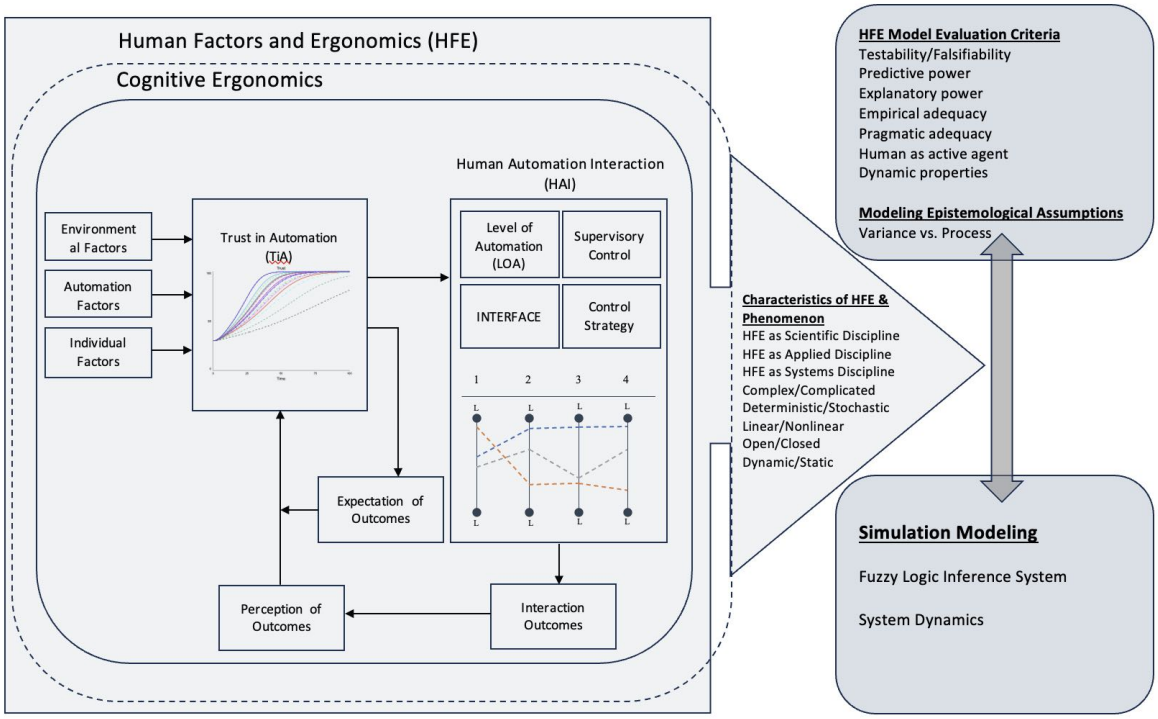


Figure 1, Overview of the dissertation research focus.

Under this understanding, and to delve deeper, the following research questions have been established:

RQ1: What constitutes the essential criteria for evaluating models within the domain of Human Factors and Ergonomics (HFE) research?

RQ2: What is the current state of Trust in Automation (TiA) models according to the criteria in RQ1?

RQ3: Are the epistemological assumptions in different modeling approaches appropriate for studying human-automation interactions (HAI)?

RQ4: How to effectively model Levels of Automation (LOAs) for Maritime Autonomous Surface Ship (MASS)?

RQ5: How can Trust in Automation (TiA) be dynamically modeled based on its internal structures?

1.4 Structure of the Dissertation

Chapter 1 sets the stage for the investigation by expressing the overall context of the dissertation, research problems, objectives, and questions, as well as an overall research approach to address the questions. Chapter 2 discusses the background of the study, shedding light on the importance of the research, the emerging concept of maritime autonomy, the rationale behind Shore Control Center (SCC) and supervisory control, the importance of Human-Automation Interaction (HAI) and Trust in Automation (TiA) in effective and safe operation of future maritime operations. The chapter is underpinned by an extensive review of pertinent literature, providing a comprehensive backdrop for the themes under investigation. Chapter 3 establishes the theoretical framework for the subsequent exploration and analysis. This chapter revisits the theoretical foundation of the Human Factors and Ergonomics (HFE) discipline and discusses today's complexity of Human-Automation Interaction (HAI) research. Chapter 4 presents the research methodology and discusses modeling as a theory-building activity. The primary focus in this chapter is on simulation as a viable tool for tackling the complexity of modeling HAI. More specifically, fuzzy logic and system dynamics are considered as the two approaches utilized in this dissertation for modeling levels and trust in automation. In Chapter 5, the philosophical foundations that guide the research are thoroughly examined. This chapter offers a reflective account of the epistemological understanding of theory/model development, guiding the concept of model validity. Chapter 6 presents the summary findings of four Articles and engages in a detailed discussion aimed at addressing the research questions introduced in the opening chapter. Chapter 7 synthesizes the research findings and offers general reflections while also critically appraising the research boundaries. Chapter 8 concludes and accentuates the contributions of the study and points out potential areas for further scholarly exploration.

2 Background

Maritime transport, one of the oldest and most crucial components of global trade, stands on the brink of a transformative leap towards autonomous shipping, a shift often referred to as "Shipping 4.0." (Lambrou & Ota, 2017). Over the last three decades, the automation levels in merchant vessel operations have progressively increased. However, the current move towards autonomous vessels promises to significantly impact the transportation of goods and the future navigation and operation of ships. While the primary research emphasis to date has been on the technical realization of more autonomous systems, largely through traditional risk assessments and technical methodologies (Dreyer & Oltedal, 2019; Thieme et al., 2018), there exists a growing need to explore the implications for human operators within the future landscape of maritime transport.

2.1 Autonomous Ships and Shore Control Center (SCC)

Maritime Autonomous Surface Ships (MASSs) have emerged as a novel domain of vehicle automation in recent years, bringing forth both fresh challenges and opportunities. The early 2010s witnessed a momentous shift towards the digital evolution of the maritime industry. This shift emphasized the automated integration of real-time data into the decision-making process (Sullivan et al., 2020). A landmark initiative in autonomous shipping was the Maritime Unmanned Navigation through Intelligence in Networks (MUNIN) project, spanning 2012–2015 (Burmeister et al., 2014). In 2017, the Norwegian firms Yara and Kongsberg embarked on a venture to develop the Yara Birkeland, a self-operating cargo vessel intended to serve three ports in Southern Norway (Yara, 2018), with aspirations for complete autonomous functionality by 2022. The advent of uncrewed Maritime Autonomous Surface Ships (MASS) promises several advantages, including the expansion of operational capabilities, such as accessing challenging and remote areas, and ensuring the safety of operators by removing them from hazardous environments (Ahvenjärvi, 2016; Norris, 2018). Insights from the MUNIN project highlighted that a variety of tasks, from adjusting shipping routes due to weather conditions or potential collisions to monitoring engine conditions for failures, could be managed by automated systems on uncrewed MASS.

The evolution of Maritime Autonomous Surface Ship (MASS) operations is set to focus on minimizing onboard crew numbers while enhancing land-based coordination and control mechanisms. This strategic shift introduces the Shore Control Center (SCC) as a primary solution. The SCC addresses the emerging requirement for centralized supervision, encompassing monitoring, and intervention tasks across MASS fleet operations. The Shore Control Center (SCC) plays a key role in supervising the operations of one or several autonomous ships from a remote location, enabling intervention in their navigation when required. An example of such a Shore Control Center (SCC) is depicted in Figure 2. The term "autonomous" in this context does not imply complete independence of the vessels; instead, it refers to a spectrum of autonomy within the ship's control system that falls short of full autonomy. According to the International Maritime Organization's (IMO) classification (IMO, 2018), at automation Levels 2 and 3, the level of automation onboard is insufficient for the vessels to navigate without human supervision. Consequently, the necessity arises for these vessels to be monitored and, when needed, remotely controlled, ensuring safe and efficient maritime operations. In future Shore Control Centers (SCC), the interaction between humans

and machines is anticipated to be the most critical element. Establishing a strong connection between the operator and the automation system becomes an important aspect of safe and effective operation. Fundamental to this interaction is the prerequisite of trust in both the system and the automation itself (Dybvik et al., 2020). Trust in Automation (TiA) will be elaborated in more detail further in this dissertation.

Figure 2, A concept of Shore Control Center (SCC) at University of South-Eastern Norway (USN) research park, Photo taken by Mehdi Poornikoo, all rights reserved.

A general conceptual framework illustrating the interaction between a Shore Control Center (SCC) operator and Maritime Autonomous Surface Ships (MASS) within a highly automated maritime environment is presented in Figure 3. The framework is structured into several layers that collectively represent the cognitive and operational dynamics of supervising autonomous vessels. At the top of the framework, the operator's cognitive processes indicate the central role of human cognition in monitoring and decision-making processes. This emphasizes that the operator's cognitive abilities, such as perception, attention, and problem-solving, are crucial in managing the operation of autonomous vessels. Directly below the operator, various automation levels specify the spectrum of automation within MASS operations. This spectrum ranges from fully manual control to semi-autonomous systems requiring significant human input, and to fully autonomous operations where human intervention is minimal. The Level of Automation (LOA) influences how the operator interacts with the system and the extent to which they need to monitor and make decisions. The concept of Level of Automation (LOA) will be further expanded in the subsequent sections. The information necessary for effective vessel supervision and control is showcased through an interface setup consisting of multiple screens. The operator interacts with the MASS, and accesses real-time data, navigational charts,

RADAR information, system statuses, and other critical operational information via the interface displays. The interface also serves as the primary tool for the operator to maintain situational awareness and execute control commands when necessary. Communication channels (e.g., satellite, 5G network) between the SCC and the autonomous vessels are crucial for transmitting commands, receiving updates, and ensuring that the vessels operate according to plan which enables continuous and reliable data exchange between the SCC and MASS. Lastly, tasks and operations involve the capabilities of the vessels and the tasks they perform independently, or guided by the supervision and intermittent interventions from the SCC. In task execution and operational capability, autonomous maritime systems are distinguished by their integration of 'perception and control' components. Perception components encompass ship positioning systems, RADAR, and additional sensors that survey the maritime environment. Control components involve mechanisms such as propulsion and steering systems. Specifically, ship maneuvering is significantly enhanced by azimuth thrusters, which combine an engine (often electric) with a propeller in a pod below the waterline. These thrusters can rotate freely by 360 degrees, allowing the ships to navigate narrow ports efficiently and safely (Thombre et al., 2020). Furthermore, the integration of Global Navigation Satellite System (GNSS) positioning with control systems into what is known as Dynamic Positioning (DP) systems enables vessels to neutralize environmental forces. This technology allows for the maintenance of precise positioning and heading, enabling ships to remain at or return to their operational stance without anchorage, or to maintain a steady course against the challenges posed by wind and waves. Despite the advancement of control systems, integrated perception systems suitable for autonomous maritime operations are still in the developmental phase (Thombre et al., 2020). Hence, a need for human operators to supervise and share control of the vessel becomes fundamental. That said, the primary focus in this dissertation is Human-Automation Interactions (HAI) which mainly involves the first three layers in the conceptual framework (Figure 3), including the operator's cognitive factors (here, Trust in Automation (TiA)), the level of automation, and the status of the automation and environmental factors received and perceived via interface displays.

Maritime Autonomous Surface Ships (MASS) may operate across a spectrum of automation levels, ranging from direct human control to fully autonomous operations where human intervention is presumably unnecessary. However, as discussed earlier, it is anticipated that a human operator will oversee the operations of uncrewed MASS, even at higher automation levels, ensuring a fallback mechanism in scenarios where the automated systems encounter difficulties or unpredictable situations (Abilio Ramos et al., 2019; Dybvik et al., 2020; Porathe et al., 2018; Størkersen, 2021). The human operator's role transition from direct engagement to supervisory control is expected to introduce new challenges (Mallam et al., 2020). In other words, the shift from Human-In-The-Loop (HITL) configurations, where human operators directly input commands and make decisions, to Human-On-The-Loop (HOTL) systems, where the human role is primarily to monitor the automated processes, may potentially exacerbate issues stemming from human operators' propensity for suboptimal monitoring (Nahavandi, 2017; Parasuraman & Riley, 1997).

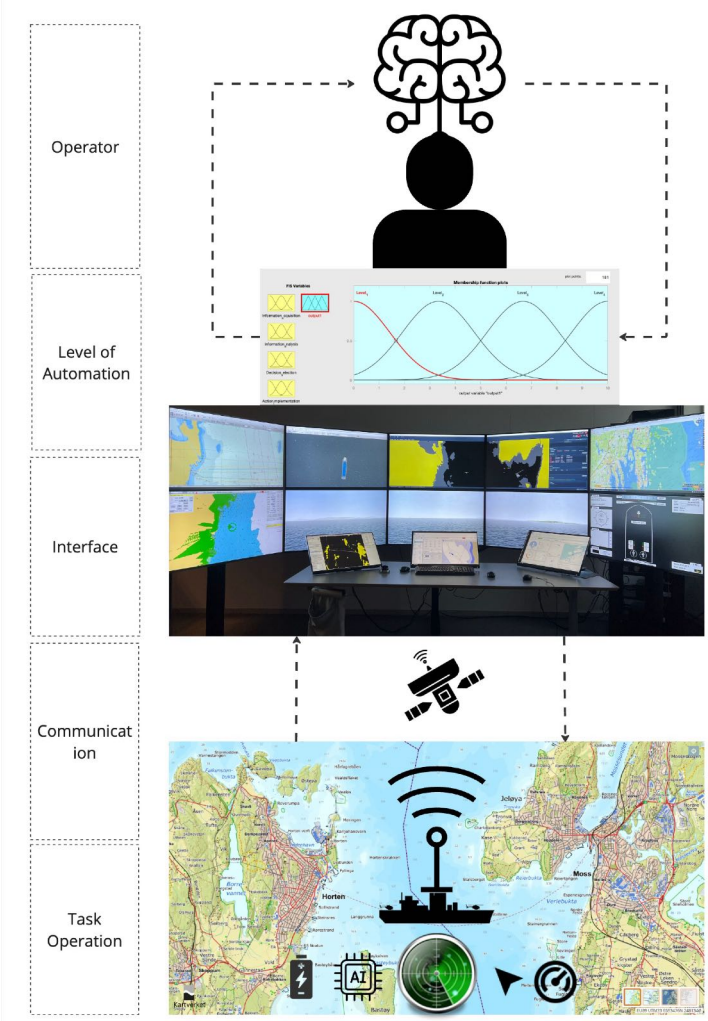


Figure 3, A conceptual framework of SCC operator and MASS operation.

The risk of monitoring lapses, compounded by an overreliance on automation, can lead to decision-making errors, resulting in incidents or accidents (Parasuraman & Riley, 1997). To mitigate these risks, it has been proposed that keeping human operators engaged through active monitoring of ship performance, weather conditions, engine functionality, and communication systems could help maintain operator alertness and early problem detection (Porathe et al., 2020). Nonetheless, the challenge remains that prolonged periods of uneventful and ordinary operation, especially in open waters, might lead to operator disengagement and a passive monitoring stance. This disengagement poses significant risks, as re-engagement or 're-looping' of the human operator during emergencies could be delayed, critically affecting response times and operational safety (Parasuraman, 2000).

With this maritime background in mind, the next section reviews the general concepts of autonomy, automation, and Human-Automation Interaction (HAI).

2.2 Autonomy, Automation, and Human-Automation Interaction (HAI)

The terms autonomy and automation are often used interchangeably (Relling et al., 2018). Autonomy, often associated with the notion of free will, manifests differently across various disciplines. In psychological terms, autonomy refers to an individual's capacity for self-governance and decision-making free from external influence (Wellman et al., 1992). In an engineering context, *artificial* autonomy implies providing machines and technologies with self-directed operational capabilities (Ziemke, 2008). The application of autonomy within engineering and automation technology typically revolves around the ability of systems to execute tasks independently of human intervention. Beer et al. (2014) propose a definition of autonomy as the degree to which a system is capable of conducting its processes and operations autonomously, without the need for external control. This general definition can encompass the autonomous functions of both biological entities, such as humans, and non-biological systems, including robots and machines (Albus & Antsaklis, 1998). Within the context of robotics, autonomy is the measure of a robot's ability to perceive its environment, formulate plans based on this perception, and execute actions to achieve specific objectives that are either assigned or self-generated, all while operating independently of external control (Beer et al., 2014).

Automation on the other hand, a term introduced in the 1950s (Diebold, 1952), is often defined as a "*device or system that accomplishes (partially or fully) a function that was previously, or conceivably could be, carried out (partially or fully) by a human operator*" (Parasuraman et al., 2000, p. 238). This definition encompasses automation across a spectrum of domains, from managing advanced cockpit systems to operating simple devices such as an automated coffee machine. The difference between automation and a machine is that automation involves executing functions that may also be performed by humans, whereas a complete and enduring transfer of a function to a machine is characterized as a machine operation (Parasuraman & Riley, 1997). With technological advancements, tasks that once necessitated human intervention, such as starting a vehicle or activating its Anti-lock Braking System (ABS), have now become standard machine operations managed by the machine itself (Adams et al., 2003).

While automation is fundamentally designed to facilitate human activities, humans typically maintain a role within the broader system. These collaborative entities, where humans and automation work together, are referred to as joint human-automation systems or human-computer systems (Johannsen, 1997), representing Licklider's (1960) vision of human-machine symbiosis, with practical applications in complex environments such as air traffic control and warehouse management (McBride et al., 2011; Rovira & Parasuraman, 2010). Human Factors and Ergonomics (HFE) research is dedicated to analyzing and enhancing the interplay and interactions between humans and automation, with a particular focus on task allocation and determining various degrees of automation.

2.3 Level of Automation (LOA)

Understanding the multifaceted interactions between humans and automation systems necessitates a theoretical framework, often facilitated by the development of models and

taxonomies. Fitts' (1951) seminal work was among the earliest to address the distribution of functions between humans and machines, advocating for a system-oriented approach where tasks are allocated based on the comparative strengths of human and machine capabilities. However, this early taxonomy did not fully embrace the dynamics of interaction and shared control, where tasks could be collaboratively managed or alternated between humans and machines based on situational demands. Various frameworks and systems for classifying levels of automation (LOA) have been introduced over the years. Sheridan and Verplank (1978) laid the groundwork by devising a 10-point scale that defined higher and lower LOAs based on the extent of autonomy. Their approach detailed the division of tasks and feedback communication between humans and automated systems, though it primarily focused on decision-making and action execution stages, somewhat overlooking the distinctions between the initial stages of information acquisition. In response to this, Endsley and Kaber (1999) refined this model to incorporate a more detailed examination of how automated systems gather and process information before making decisions. Their revision introduced a structured categorization of automation functions into four main activities: monitoring, generating, selecting, and implementing, thereby offering a more developed view of the automation process.

Parasuraman et al. (2000) developed a model that echoed Endsley and Kaber's emphasis on the varying degrees of automation. Their framework provided a structured approach to categorize human-automation interactions based on a multi-stage model of human information processing. This taxonomy identifies four primary *types* of automation, relating to different stages of information processing: (1) information acquisition, where automation assists in filtering and focusing attention on relevant external information, (2) information analysis, where it helps in integrating and interpreting information, (3) decision and action selection, where automation contributes to determining and selecting appropriate actions based on analyzed information, and (4) action implementation, where it executes the chosen actions. Each type encompasses a spectrum of automation *levels*, from entirely manual to fully automated processes, as illustrated in Figure 4.

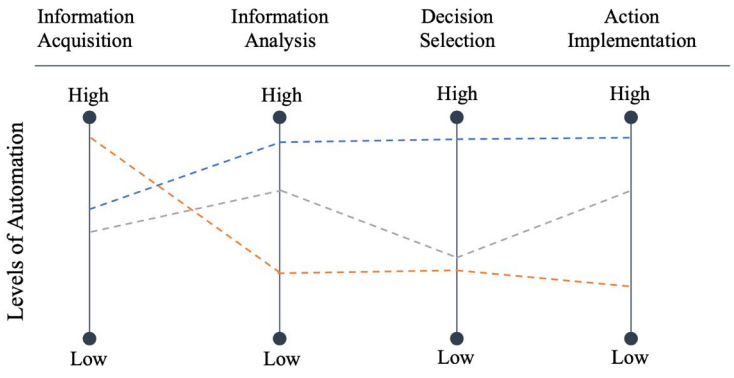


Figure 4, Systems with different types and levels of automation, adapted from Parasuraman et al. (2000)

In this model, the initial stage (i.e., information acquisition) plays a focal role in capturing and processing sensory input. This phase augments human sensory abilities, facilitating the monitoring of various environmental parameters. It encompasses technologies that gather environmental data, such as radar systems and thermal imaging devices. When automation

reaches advanced stages of information acquisition, it has the capability to sort and prioritize this sensory data, similar to how a Vessel Traffic System (VTS) or Air Traffic Management System (ATMS) sequences vessels and aircraft (Sheridan et al., 2002). Subsequently, the information analysis phase of automation undertakes functions that parallel human cognitive activities, notably those associated with working memory. In this phase, the automation system might produce predictions, merge diverse data inputs, or condense information for presentation to the user. Distinct from the acquisition phase, this analysis phase actively interprets and processes the data. In the decision-selection phase, automation assumes the role of choosing between different decision-making alternatives. Such systems may, for instance, determine optimal flight paths for aircraft to evade bad weather (Ng et al., 2009; Xie & Zhong, 2016), route planning under different tidal conditions in maritime navigation (Pan et al., 2021), or assist medical professionals by suggesting possible diagnoses (Thanh et al., 2017). The final phase, action implementation, involves automation executing the selected decisions. This could entail completing an entire task or its constituent parts, such as the autopilot feature in ship operation.

Parasuraman et al. (2000) further recommended a series of iterative procedures (Figure 5) to utilize the proposed framework for automation design. The framework aimed to determine the degree to which tasks should be automated, considering the effects on human operators and the automation system itself. The process begins by selecting a preliminary level of automation for each category, which is then assessed using primary evaluative criteria related to human performance outcomes. If required, adjustments are made to the automation level based on this assessment. For example, a decision support system at a fundamental level of automation would be assessed against primary criteria such as human workload and situational awareness, which might necessitate adjustments to the automation level to mitigate workload. Following this, secondary criteria such as the dependability of automation and the implications of decision outcomes are assessed, potentially leading to further calibration of automation levels. This evaluative cycle is systematically applied to each category of automation, ideally ensuring that the end result is a harmonized blend of human and automated processes. Furthermore, it facilitates the identification and resolution of design challenges by determining the optimal levels or spectrum of automation.

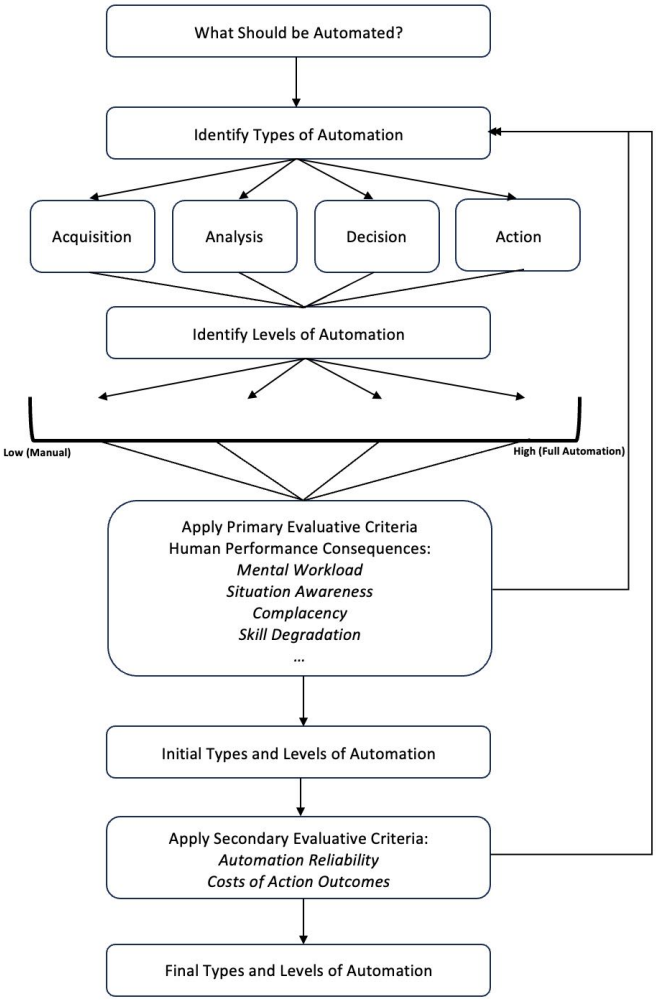


Figure 5, Application of levels and types of automation, adapted from Parasuraman et al. (2000)

Today, the use of taxonomies for levels of automation (LOAs) is prevalent in scholarly literature, primarily for their ease of understanding and as a foundation for further design considerations. These taxonomies have recognized the varied nature of automation support, yet their application across different domains requires more specificity and clarity (Poornikoo & Øvergård, 2022). Recently, Vagia et al. (2016) reviewed the existing LOA taxonomies and reported a lack of consensus on the definitions and levels of automation. Originally devised for task-oriented and functional operations (Endsley & Kaber, 1999; Endsley & Kiris, 1995; Riley, 1989; Sheridan & Verplank, 1978), the taxonomies are now employed by policy organizations and classification bodies to categorize end products of manufacturing processes. For example, the taxonomy by the Society of Automated Engineers (SAE, 2014) delineates five automation levels for vehicle driving modes, focusing on operational and tactical tasks such as steering and environment monitoring. This taxonomy aims to support manufacturing technology classifications, future design initiatives, and inform road traffic regulations. Similarly, the maritime sector has developed several taxonomies for maritime autonomous surface ships (MASS) as shown in Figure 6, indicating degrees of autonomy and categorizing autonomous vessels not merely by automated tasks but as advanced, complete systems. This has resulted in the application of LOAs becoming quite context-specific, with interpretations varying based on

the industry rules and standards, as well as the analytical level. Furthermore, the prevailing levels of automation for MASS primarily focus on navigational functions as the benchmark for determining the ship's level of autonomy. However, according to the International Maritime Organization's (IMO) description of MASS, an autonomous ship is characterized by its capacity to function to varying extents without human intervention (IMO, 2018). This definition implies that an autonomous ship's functionality extends beyond mere navigation to include a wide range of critical operations such as maintenance, cargo management, anchoring, etc. Poornikoo and Øvergård (2022) outlined the general limitations of the existing LOA taxonomies from an operational perspective and offered a simulation-based model for LOA.

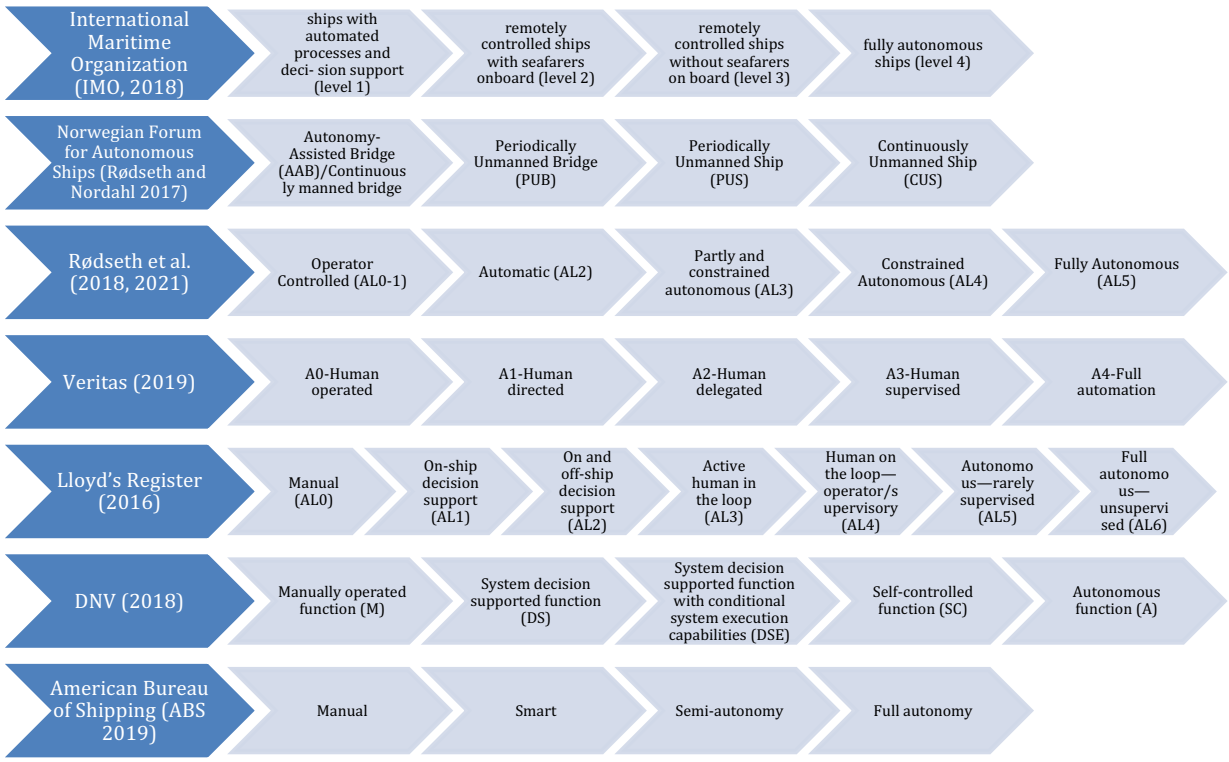


Figure 6, Levels of Automation/Autonomy for MASS

2.4 Supervisory Control

For a long time, the addition of automation in systems was viewed as simply replacing human tasks with machine operations, a concept referred to as the substitution myth (Woods & Dekker, 2000). However, this view is a limited and inadequate representation of the true impact of automation. Automation technology significantly transforms human practices, compelling individuals to adjust their skills and routines (Dekker & Woods, 2002). Particularly noteworthy is the shift in the role of the human operator. Given the current state of technology, automation lacks the capacity for "intelligent" adaptability in unforeseen situations, which necessitates human supervision and, at times, direct intervention. The operator now primarily monitors the system's actions, comprehends these actions, looks out for deviations and failures, and intervenes when necessary (Moray et al., 1986; Sheridan, 2017, 2021; Sheridan et al., 1978).

This shift features an irony; while designers aim to reduce the operator's active role, they still rely on the operator for tasks that cannot be automated (Bainbridge, 1983).

The concept of human supervisory control is similar to a human supervisor interacting with subordinates. Just as a supervisor issues instructions that subordinates interpret and execute, translating complex information into actionable tasks, automation enables similar dynamics between humans and machines (Sheridan, 2021). The level of automation delegated by the supervisor to their subordinates—or in this case, to automated subsystems—is influenced by the perceived intelligence of those executing the tasks, affecting both the depth and the duration of the commands given. This form of supervisory control spans a wide array of applications, including the management of vehicles such as aircraft, spacecraft, and maritime vessels, control of continuous processes including oil, chemical, and power generation industries, and supervision of robots and discrete manufacturing tasks. It also extends to medical systems, home automation technologies (e.g., heating and appliance management), and various other domains where human-machine interaction is focal. At its core, human supervisory control, sometimes simply referred to as "supervisory control", involves human operators intermittently setting goals for computers that manage internal control loops via electromechanical actuators, tasks, and feedback sensors (Sheridan, 2021). In a more expansive view, supervisory control encompasses any interaction with a computer interface that modifies data or generates control actions as shown in Figure 7.

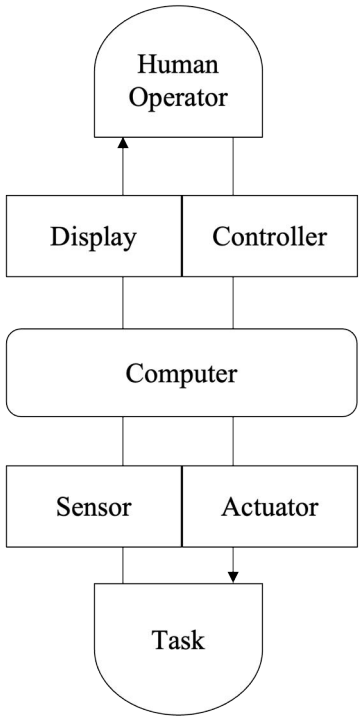


Figure 7, Concept of Supervisory Control (Sheridan, 2021)

The roles of a human supervisor in supervisory control systems encompass several key functions including: (1) offline planning of tasks and procedures, (2) instructing the computer based on the plan, (3) online monitoring of automated actions to ensure alignment with the plan and to identify any failures, (4) intervening either post-achievement of goals or during

emergencies to establish new objectives and procedures, and (5) learning from these experiences to improve future task execution. These roles typically follow a sequential timeline within the performance of a task (Sheridan, 2021), and may also incorporate varying levels of computer assistance (i.e., LOAs) in both information acquisition and control execution.

2.5 Complacency and Automation Bias

While automation holds significant promise, it also poses the risk of negatively altering operator behavior (Parasuraman & Riley, 1997). Two notable issues, complacency, and automation bias are extensively examined in human-automation interaction (HAI) research. Complacency arises from excessive reliance on automation, leading to insufficient verification of system states or the underlying data. Consequently, human performance suffers when automation fails (Wickens et al., 2015). This concern is particularly relevant to highly reliable yet imperfect automation systems. For instance, Bagheri and Jamieson (2004) found that in multitasking environments, operators were less adept at detecting system malfunctions when automation was deemed highly reliable. Operators may preferentially attend to tasks not supported by automation, neglecting those aided by it, despite competing demands for their attention. Although this strategy may seem logical (Moray & Inagaki, 1999), it has been shown to lead to miss of critical safety information (Metzger & Parasuraman, 2005).

Automation bias manifests when operators depend on incorrect guidance from automated aids, leading to a decline in decision-making performance (Wickens et al., 2015). Automated alert systems, such as collision warnings in maritime or the Traffic Collision Avoidance System (TCAS) in aviation, are designed to enhance human cognitive processes during high-risk situations by issuing alerts or recommendations. However, despite the fallibility of such systems, operators frequently evade full data analysis, resulting in less-than-optimal decisions (Mosier et al., 1998; Mosier & Skitka, 1999; Rovira et al., 2007).

Automation bias and complacency play a critical role at each stage of information processing and levels of automation. They manifest as insufficient system status monitoring or an over-reliance on automation, overlooking additional information that could influence decision accuracy (Yamani & Horrey, 2018). Rovira et al. (2007) observed that as the level of automation increased (i.e., from providing a list of possible options to recommending the singular best action), operators' decision-making accuracy deteriorated. These issues suggest a direct relationship between the complacency and automation bias induced by automation and the ineffective strategy of operators in allocating attention and prioritizing tasks. In a similar fashion, the remote supervisory control of MASS may potentially introduce new challenges, as humans can be relatively ineffective at sustained monitoring tasks (Nahavandi, 2017; Parasuraman & Riley, 1997). Subpar monitoring of automated systems can result from an operator's excessive reliance on the system, which can lead to decision errors and consequently, incidents and accidents (Parasuraman & Riley, 1997). Empirical studies suggest that behavioral reliance on automation is strongly influenced by the level of trust that users have in the automation system. Trust shapes the human operator's readiness to delegate responsibilities to automated systems and determines the extent and frequency of interventions in automated

processes. Trust also guides the operator's acceptance of suggestions made by the automation (Lee & See, 2004). Operators tend to depend more on automation systems that they trust to be reliable and effective (De Vries et al., 2003; Lee & Moray, 1992; Merritt, 2011; Merritt & Ilgen, 2008; Wang et al., 2009). Consequently, the reliance behavior that the human supervisor adopts has a profound impact on both mission outcomes and overall system performance (Clare, 2013; Gao et al., 2013). Mallam et al. (2020) conducted interviews with maritime subject matter experts, revealing that trust is a predominant theme when assessing the potential impact of autonomous shipping. Considering the significant role that Trust in Automation (TiA) plays in the performance of joint human-automation systems, the next section reviews the implication of TiA in human-automation interaction (HAI), its facets, dynamics, and models.

2.6 Trust in Automation (TiA)

Trust is widely regarded as a psychological construct, related to expectation and anticipation of reliable actions of another party (de Vries, 2005). Trust is also perceived as a multi-faceted and dynamic phenomenon (Atoyan et al., 2006; Dzindolet et al., 2003).

Trust in Automation (TiA) originated from early theories that drew parallels with the psychological understanding of interpersonal trust (Muir, 1994a; Muir & Moray, 1996). Similar to interpersonal trust, TiA contains a sense of risk or vulnerability from the trustor's perspective, demanding a foundational level of trust for its development (Corritore et al., 2003; Evans et al., 2011; Evans & Krueger, 2011; Lee & Moray, 1994; Lee & See, 2004). This is because, in dynamic and urgent scenarios, individuals interacting with automated systems may struggle to perceive and analyze all essential details necessary for effective situation management (Moray et al., 2000). Under such circumstances, they must operate under risk and uncertainty, lacking comprehensive knowledge of all relevant aspects to accurately assess the situation (Rajaonah et al., 2006). Despite debates on the similarities and differences between interpersonal trust and TiA (Madhavan & Wiegmann, 2007), the importance of Trust in Automation (TiA) has become a central point in Human-Automation Interaction (HAI) research to bolster joint system performance (Lee & Moray, 1992; Lee & See, 2004; Muir, 1994a; Muir & Moray, 1996).

Various definitions of TiA (e.g., Table 1) converge on the premise that TiA embodies a belief, attitude, or expectation in the automation's capability to fulfill its intended task. Thus, TiA emerges through a continuous process of aligning expectations with actual observations of automation performance, especially in contexts where the user bears a significant risk (Kenesei et al., 2022; Li et al., 2019; Muir, 1994a; Sheridan & Hennessy, 1984).

Table 1, Definitions of Trust in Automation (TiA)

			Parties involved	
	Definition	**Nature of trust**	**Trustor**	**Trustee**
Muir (1987)	"...the intervening variable [between the automation and the supervisor's responses to the automation] that mediates supervisors' intervention behavior."	Intervening variable	Supervisor	Automation

Mayer, Davis, and Schoorman (1995, p.712)	"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 other party".	Expectation	Trustor	Trustee
Moray and Inagaki (2000)	"...an attitude which includes the belief that the collaborator will perform as expected, and can, within the limits of its designers' intentions, be relied on to achieve the design goals"	Attitude, Belief	Human	Collaborator
Madsen and Gregor (2000)	"the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent agent."	Confidence	User	Intelligent Agent
De Vries (2005)	"...the expectation of a user about the system, that the system will perform a certain task for him or her, while the outcome of that task is uncertain, in that it can have both positive and negative consequences"	Expectation	User	System
Lee and See (2004)	"the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability"	Attitude	Individual	Agent
Biros et al. (2004)	"...having confidence in and entrusting the system automation to do the appropriate action"	Confidence	Human	System Automation
Rajaonah et al. (2006)	"...a psychological state resulting from knowledge, beliefs, and assessments related to the decision-making situation, that creates confident expectations for human machine system performance and guides operator reliance on automation "	Psychological state based on knowledge, beliefs and assessments	Operator	Automation
Madhavan and Wiegmann (2007)	"...the expectation of, or confidence in, another and is based on the probability that one party attaches to co-operative or favorable behavior by other parties "	Expectation Confidence	One party	Another party

Applied to Maritime Autonomous Surface Ships (MASS), trust becomes a critical factor as the operator transitions from an active contributor in direct navigation and control to a more passive monitoring role. While the operator can, in theory, always intervene and undertake manual control, the fundamental principle of autonomous maritime operations does not mandate such interventions during standard operations. When a system limitation is detected, the operator is alerted through a clear notification, compelling them to manually take over the navigation and control of the ship within a predetermined period. In such a scenario, trust serves as a mediating factor in the interplay between the automation and its use. This mediating variable can be studied from two perspectives: Trust as a consequence of various influences (trust factors), and trust as a precursor that shapes subsequent reliance on the automation (Masalonis & Parasuraman, 1999).

2.6.1 Trust as Consequence (Trust Factors)

Trust can be influenced by a variety of factors, generally categorized into elements related to the automation, the operator, and the environment. Comprehensive discussions of these factors can be found in the studies and meta-analyses by Merritt and Ilgen (2008), Schaefer et al. (2016), Schaefer et al. (2014)), Hancock et al. (2011), Lee and See (Lee & See, 2004), and Hoff and Bashir (2015). An overview of the prominent factors is discussed in the following section.

2.6.1.1 Automation-related factors

Trust is significantly influenced by the automation's characteristics. Some of the main characteristics include:

A. Automation performance (reliability)

The level of trust users place in automation is significantly influenced by the automation's performance. A reliable automation fosters greater trust and is subsequently more frequently utilized (Muir, 1994a; Muir & Moray, 1996). Automation reliability, also referred to as system competence (Muir, 1987), is characterized by the system's consistent performance. This performance level directly correlates with the trust users place in the system, to the extent that users may prefer automation over their own capabilities for operating a system (Merritt et al., 2013). The expectation and actual reliability of a system are key in forming Trust in Automation (TiA) (de Vries, 2005; Kazi et al., 2007; Moray et al., 2000). The consistency of system reliability plays an important role in its predictability and perceived trustworthiness; systems that demonstrate stable reliability are deemed more predictable and, therefore, more trustworthy (Muir & Moray, 1996; Parasuraman et al., 1993). Additionally, trust is subject to the primacy-recency effect, where initial low reliability can lead to a long-term lack of trust and usage (Atoyan et al., 2006).

B. Automation predictability and transparency

The predictability of a system is closely linked to its perceived reliability and the consistency of its performance. The expectation of a system's predictability, as suggested by Muir (1987), is considered an essential factor impacting trust. Predictability is intertwined with the system's transparency; because a system's actions and intentions must be understandable and logically explicable to users for its behavior to be anticipated. Ososky et al. (2014) define system transparency as the extent to which the actions and intentions of a system are visible and comprehensible to human operators. Achieving this may involve designing automation that behaves in a manner akin to human decision-making or creating systems capable of articulating their processes and reasoning to users (Ghazizadeh et al., 2012; Sarter & Woods, 1997; Seppelt & Lee, 2019). The lack of feedback and transparency on automated processes is a common root cause of automation-related accidents (Norman, 1989). Endsley & Kiris (1995) articulate the challenge for designers in providing sufficient feedback to keep the operator informed without causing information overload. Transparency in automation, therefore, is crucial, enabling operators to understand the automation's functioning or failure. Simpson et al. (1995) argue that trust in a system is contingent upon its ability to demonstrate competent performance and enable predictions of its reliability. Transparency also facilitates the formation and updating of mental models about the system (Matthews et al., 2020; Miller, 2021), preventing unexpected

automation behaviors and clarifying the system's limitations or errors. The ability of automation to self-explain, especially during errors, enhances trust and reliance (Dzindolet et al., 2003).

C. System malfunction (Faults/errors/failures)

Faults or system malfunctions tend to erode trust, with the effect varying based on the failure's severity and frequency. The larger the fault, the more significant the reduction in trust (Moray 1992). However, if automation provides a suboptimal performance that does not result in an absolute and significant fault, this may reduce the negative impact on trust, as observed in Lee and Moray's (1994) study where a constant fault resulted in increased TiA as participants accustomed to the fault and established compensation strategies. The type of fault also plays a role, with false alarms primarily affecting operator compliance and 'misses' impacting reliance (Dixon et al., 2007). Trust typically experiences a significant decline following errors and recovers slowly, even if system performance is promptly restored (Lee & Moray, 1992). The consequence of a fault may be just as crucial as its magnitude in affecting trust. Masalonis et al. (1998) found that trust was lower when an automated aid failed to notify supervisors about a possible aircraft encounter compared to when it issued a false alert.

The literature indicates that foreknowledge of potential system faults can mitigate their negative impact on trust, suggesting the importance of system transparency over its actual performance (Beggiato & Krems, 2013; Dzindolet et al., 2003; Riley, 1996). A predictable system, even with ongoing minor errors, can still be used and trusted if users understand its limitations and behavior (Lee & Moray, 1992; Ma, 2005; Muir & Moray, 1996). However, discrepancies between users' expectations and system performance can negatively affect trust, even if the automation operates as designed (Lee & See, 2004). Madhavan et al. (2006) found that the perceived difficulty of a task could also influence trust in automation, with failures in seemingly simple tasks being particularly damaging to trust.

D. Level of Automation (LOA)

The level of automation which spans from minimal assistance to complete autonomous, substantially affects the user's trust. Walliser (2011) noticed that the automation level impacts the operator's trust calibration, as well as their performance during system errors. Higher LOA have been associated with longer response times to system failures compared to lower levels (Niederée et al., 2012a, 2012b; Shen & Neyens, 2014). While lower LOA necessitates user vigilance for system errors, higher levels, though less prone to expected errors, may still surprise users with unexpected behaviors, potentially undermining trust.

2.6.1.2 Operator-Related Factors

Trust in automation extends beyond the technical features of the system and is profoundly shaped by an individual's subjective interpretation of these characteristics (Lee & See, 2004). Merritt and Ilgen (2008) found that a person's perception of automation is shaped by both the actual features of the automation and their natural tendency to place trust (trust propensity) in automation.

A. Demographics

In terms of demographic attributes, culture, age, gender, and personality have been identified as key factors influencing trust (Hoff & Bashir, 2015). Research has demonstrated variations in the interaction and trust levels with automation across different cultures (Heimgärtner, 2007; Hoff & Bashir, 2015). Sanchez et al. (2014) suggest that, at the onset of interaction with an automated system, older individuals tend to exhibit lower reliance on the automation, aligning their trust levels more closely with the system's reliability changes. However, a recent study by Hartwich et al. (2019) explored the relationship between age groups and trust in automated driving systems, revealing no notable variances across different age groups. Similarly, the influence of gender on trust in automated systems has not yet reached a conclusive agreement in the literature (Hoff & Bashir, 2015).

B. Personality traits

In the context of personality traits, the Five-Factor Model of Personality (John & Srivastava, 1999; McCrae & John, 1992) has been frequently used in examining how general traits relate to trust. Specifically, extraversion is reported to be positively correlated with higher levels of interpersonal trust (Evans & Revelle, 2008), a trend that extends to trust in automated systems (Merritt & Ilgen, 2008). In contrast, neuroticism typically shows a negative correlation with interpersonal trust and may influence skepticism towards automation, as inferred from studies on acceptance of automated recommendations (Szalma & Taylor, 2011). Further, traits such as agreeableness and conscientiousness have been found to positively affect the initial trust individuals place in automation (Chien et al., 2016), suggesting that these personality traits may significantly influence one's propensity to trust automated technologies.

C. Experience

Experience with automated systems can influence trust, as individuals form expectations about these systems based on their observed reliability. Muir (1994) posits that expert operators, familiar with system intricacies, are less likely to exhibit confirmation bias compared to novice users. However, Riley (1994) found that automation experience did not significantly alter the relationship between workload, automation reliability, and usage, suggesting that further investigation is needed, particularly in high-fidelity simulations and in testing the hypothesis on the persistence of belief relative to experience. Research by Sanchez et al. (2014) demonstrated that the effect of low system reliability on trust varies according to the user's level of familiarity with the system. Additionally, findings by Manzey et al. (2012) suggest that negative experiences with an automated system have a more profound impact on trust than positive experiences. Therefore, both the quantity and quality of interactions with a system play vital roles in shaping the degree of trust and reliance placed on it.

2.6.1.3 Environmental-Related Factors

Environmental factors influence the relationship between trust and interaction with automated systems, though they may not directly impact trust. Hoff and Bashir (2015) highlighted that the unfamiliarity of a situation, along with the degree of autonomy afforded to the operator, and the operator's capacity to evaluate automated versus manual execution can impact the relationship between trust and reliance on automation. In scenarios where individuals have the opportunity to assess and confirm the automation's accuracy, trust is more likely to dictate

reliance on the automation. Moreover, the perceived advantages and potential risks associated with employing automation, alongside task requirements and the operator's workload, play important roles in shaping this dynamic.

2.6.2 Trust as Precursor (Trust outcomes): Automation Misuse, Disuse, and Abuse

Automation misuse and disuse, examined by Parasuraman and Riley (1997), encapsulate common trust outcomes in human-automation interaction. Misuse or "*overreliance on automation*" (Parasuraman and Riley 1997, p. 230) is characterized by an uncritical reliance on automation, often leading to its overuse, and stems from two main factors: automation bias and complacency. Both factors contribute to insufficient monitoring due to diminished human engagement (Parasuraman & Manzey, 2010; Parasuraman & Riley, 1997). Automation bias, the predisposition to accept automated feedback as accurate, emerges from the human inclination towards minimizing cognitive effort, thereby preferring to trust automation's correctness (Dzindolet, Beck, et al., 2001; Dzindolet, Pierce, et al., 2001; Goddard et al., 2014; Mosier et al., 1998; Skitka et al., 1999, 2000; Wang et al., 2008). Complacency manifests when monitoring is suboptimal, adversely affecting the performance of the joint system. This tendency is exacerbated in high-workload and high-stakes environments, where users might opt to depend on even flawed automation (Dixon et al., 2007; Wickens & Dixon, 2007).

In contrast to misuse, disuse occurs when the automation remains underutilized despite its high reliability (Parasuraman & Riley, 1997). Automation disuse ranges from minimal use of automation to complete reliance on manual operation. It often results from discrepancies between expected and observed automation performance or when a user's confidence in their own ability to perform a task surpasses their trust in the automation's effectiveness (Lee & Moray, 1992).

When automation is employed in contexts beyond its intended design or in an inapplicable situation, this refers to automation abuse (Parasuraman and Riley 1997). Such abuse can result in system malfunctions and diminished performance of the automation. An example of this could be the activation of automated lane-keeping in an urban driving setting, which is designed specifically for highway use.

While the concepts of Trust in Automation (TiA) and its resultant outcomes are well-established, a comprehensive review of the literature alongside recent empirical findings suggests a more complex relationship between TiA and trust outcomes than previously understood. This complexity suggests that predicting the outcomes of human-automation interactions based solely on the level of TiA can be excessively simplistic. Instances have been observed where users exhibit high levels of TiA yet opt for manual operations (Lee and Moray, 1992), or users under high workload conditions have been observed to misuse automation they do not fully trust (Biros et al., 2004; Daly, 2002). Despite understanding the behavioral aspects of TiA remaining crucial (Drnec, Marathe, Lukos, et al., 2016; Drnec, Marathe, Metcalfe, et al., 2016), these findings suggest that the relationship between TiA and behavioral outcomes

(e.g., the rate of intervention or attention levels) can better be perceived as a nonlinear dynamic process.

2.6.3 Trust Development

Muir (1994) argues that TiA's development hinges on meeting three expectations during interactions with automation. These are *technical competence*, *persistence* (which might be more appropriately described as predictability), and *fiduciary responsibility*. Each of these expectations plays a varying role in the evolution of TiA over the duration of automation usage. Initially, the automation's perceived technical competence, or its ability to accurately fulfill its designed functions, may be paramount. As time progresses, the focus may shift to other aspects. Persistence, in this context, is closely tied to the automation's reliability; it is the anticipation that the automation will consistently perform in a similar manner under comparable conditions in the future. Fiduciary responsibility encompasses the user's expectation that the automation will be accountable for the tasks it is designed to perform, thereby allowing the user to allocate fewer personal resources to those tasks. The significance of these expectations in the dynamics of TiA varies at different interaction phases with the automation.

Upon initial exposure to an automated system, human users often face a scarcity of information to assess the system's trustworthiness. Early expectations of Trust in Automation (TiA) are influenced by preconceived biases towards automation and initial impressions of the system's design. These basic assessments initiate the development of TiA (Dzindolet et al., 2003; Lee & See, 2004; Merritt, 2011; C. Miller et al., 2005; Muir & Moray, 1996; Nass et al., 1996; Pak et al., 2012; Parasuraman & Miller, 2004). As users familiarize themselves with the system, they experiment with different interaction strategies which facilitates a deeper understanding of the system's capabilities. This exploration phase is critical for assessing the system's competence, a key determinant of TiA in the beginning stage. Although, humans often struggle to accurately estimate system competence due to various biases and limitations (Madhavan et al., 2006; Merritt et al., 2014; Sheridan & Hennessy, 1984; Verberne et al., 2012), once a judgment on system competence is formed—accurate or not—predictability or persistence in the system's performance becomes the key factor in sustaining TiA over time. Consistent performance, particularly with an error rate maintained at or below 30%, is generally claimed to be sufficient for users to continue relying on the system (Wickens & Dixon, 2007). TiA evolves dynamically as users accumulate experiences with the automation, influencing their interaction decisions and subsequent behaviors.

2.6.4 Models of TiA

The conceptual understanding of trust has become more cohesive over recent years, yet there remains no universally accepted definition nor model for Trust in Automation (TiA), likely due to the situational specifics inherent to trust (Kohn et al., 2021). This absence of a singular model implies that interpretations of trust may differ according to the objectives and theoretical underpinnings chosen by each researcher, provided that these are explicitly integrated within their conceptual framework.

Muir (1987) introduced an initial model of trust, integrating three dimensions of expectations based on Barber's (1983) research and three levels of experience, drawing from Rempel et al. (1985). The model defines three dimensions of expectations encompassing *persistence*, which is the belief in the stability of natural, physical, biological, and moral social orders; *technical competence*, reflecting trust in the predictable actions of another agent; and *fiduciary responsibility*, which is the anticipation that the trusted party will act in the best interest of the trustor. These facets of technical competence intersect with personal experiences across *predictability*, *dependability*, and *faith* levels, suggesting a perpendicular relationship between the dimensions of expectations and experiences (Muir, 1994). Consequently, perceptions of an automated system's *persistence*, *competence*, and *responsibility* are influenced by an individual's prior experiences with the system (predictability, dependability, and faith), shaping the trust they place in the technology. This trust may be accurately or inaccurately aligned with the system's actual attributes. According to Muir (1994), trust in automation evolves from initial reliance on the system's consistent actions to a deeper trust founded on perceived reliability after extensive interaction. The ultimate level of trust, therefore, is founded on both empirical evidence and a leap of faith beyond rational considerations (Adams et al., 2003). Further investigations confirm that significant elements of trust in automation are encapsulated within models of interpersonal trust (Muir & Moray, 1996). In a controlled setting, individuals assessed their trust in a system based on its operational performance, corroborating Muir's trust model and highlighting the nature of trust as influenced by the duration of system interaction (Muir & Moray, 1996).

Lee and Moray (1992) extended the foundational ideas proposed by Muir (1994), incorporating the constructs postulated by Barber (1983) and Rempel et al. (1985), but further enriching them with additional contextual elements specific to their research. According to Lee and Moray (1992), trust is initially grounded on basic beliefs about the nature and structure of society, forming the bedrock upon which further aspects of trust are constructed. In this model, trust is dissected into three primary components: *performance*, *process*, and *purpose*. The *Performance* dimension reflects the observed and historical characteristics of an automated system, including its reliability and predictability, essentially focusing on the outcomes of the system's actions. *Process* delves into the suitability and methodology of the system's operations, providing insight into the system's operational logic and procedural correctness. Lastly, *Purpose* considers the original intent and application for which the system was designed, underpinning the rationale behind its functionalities. In their empirical study on supervisory control, Lee and Moray (1992) observed shifts in trust and control approaches as users interacted with an automated processing unit. Their findings highlighted the significant roles played by system performance and failures in shaping subjective trust levels, suggesting that these trust facets (performance and process) significantly inform other trust dimensions such as predictability, dependability, and faith.

Lee and See (2004) compiled theories and insights from diverse fields such as interpersonal relationships, psychology, sociology, and organizational behavior to construct a comprehensive model that elucidates the evolving nature of trust in interactions with automated systems. Anchoring their discussion in the Theory of Reasoned Action (TRA) proposed by Fishbein and

Ajzen (1975), they adapt its principles to the context of TiA. The TRA postulates that human behavior is under volitional control and significantly shaped by behavioral intentions, which themselves are the result of a combination of attitudes, beliefs, and societal norms. Behavioral intention in this framework is seen as a reflection of the effort and motivation an individual is willing to invest to enact a particular behavior. In this light, beliefs are interpreted as individual perceptions regarding the likelihood that engaging with a particular object (e.g., an automated system) will result in specific outcomes. These beliefs, informed by relevant information, shape one's attitudes towards the object, encapsulating evaluative judgments along dimensions such as good-bad or pleasant-unpleasant. Unlike beliefs, which are situationally specific, attitudes are broader and more stable evaluations that transcend specific contexts.

Adapting these concepts to the domain of automated technology, Lee and See (2004) proposed that trust acts as a critical mediator in the relationship between users and automation, influencing and being influenced by the interaction dynamics. This reciprocal relationship is captured in their model (Figure 8), which outlines a closed-loop process where interaction with automation feeds into trust, which in turn affects subsequent interactions. This interactive cycle is modulated by external factors including environmental context, system characteristics, and user traits. Lee and See (2004) elaborate on the concepts of *detail* and *abstraction* as they pertain to the understanding and development of trust in automation. *Detail* pertains to the granularity of trust, which might focus on specific elements such as the operational modes of an automation system or the system as a whole. *Abstraction*, on the other hand, captures broader considerations such as the system's overall performance, its operational processes, and the objectives for which it was designed, following the earlier insights of Lee and Moray (1992).

Lee and See (2004) suggest that for trust to be accurately calibrated and deemed appropriate, both the granularity (detail) and the broader context (abstraction) of the automation's capabilities should be communicated effectively to the user. Importantly, the model underscores the impact of how information is conveyed through the user interface, suggesting that the nature and presentation of information can significantly influence trust dynamics in automated systems. This conceptual framework offers a comprehensive understanding of trust in automation, emphasizing its dynamic nature and the multifaceted influences that shape trust-based behavioral decisions in dynamic decision-making scenarios involving automated technologies. Lee and See's model made a significant contribution toward understanding the psychological underpinnings of trust formation in automation. By positing that trust is dynamic and influenced by an array of factors, the model offers an extensive perspective on trust in automated systems. Since its introduction in 2004, the model has become a cornerstone for discussions surrounding trust in automation, especially its conceptualization of trust as an attitude-based phenomenon. Despite its widespread recognition, empirical validation of the model's core principles remains limited. Moreover, the model's variables and their interrelations within the trust formation process need clearer definitions and more precise explanations to enhance empirical testing and theoretical utility.

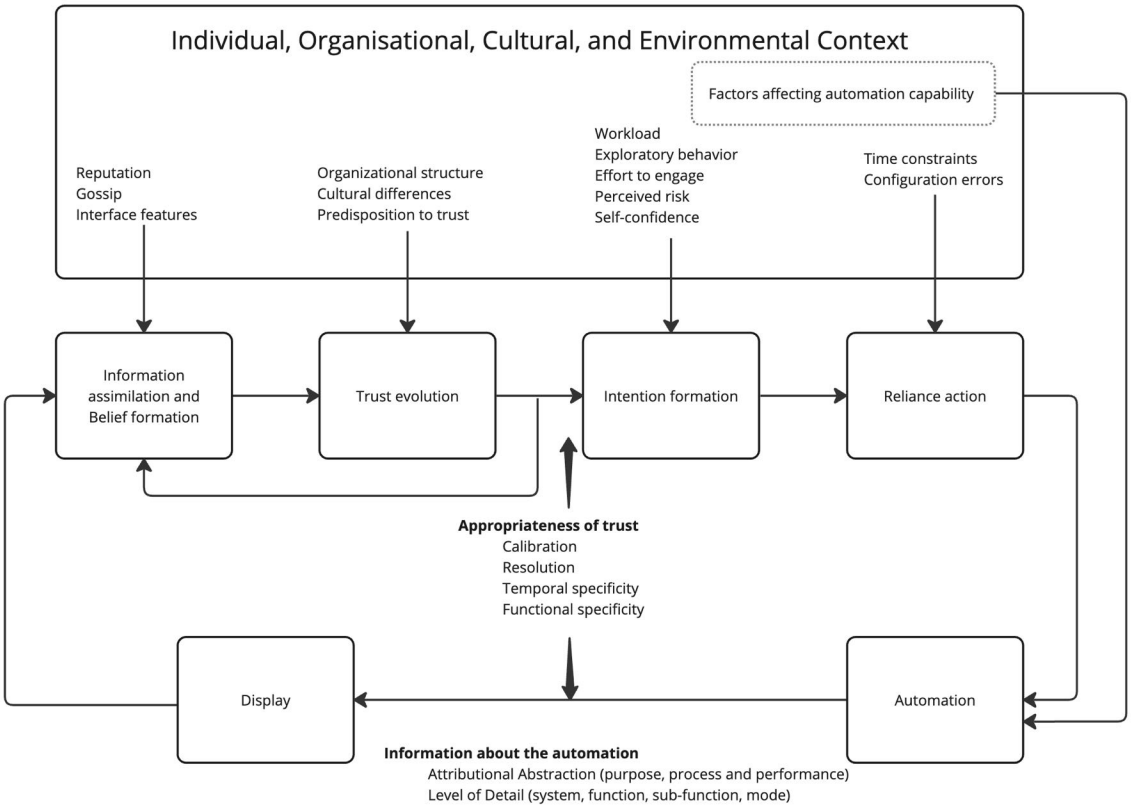


Figure 8, Trust in Automation and Reliance (Lee and See, 2004)

The development of trust in automated systems is multifaceted, encompassing a variety of human internal characteristics, system features, situational dynamics, and environmental factors (Hancock et al., 2011; Lee & See, 2004; Merritt & Ilgen, 2008). Drawing from a comprehensive review of 127 studies on human-automation interaction, Hoff and Bashir (2015) condensed these factors into a three-layered model that categorizes trust into *dispositional*, *situational*, and *learned* dimensions, as illustrated in Figure 9.

Dispositional trust encompasses an individual's inclination to trust automated systems, influenced by stable personal characteristics such as demographics, cultural background, and personality traits. This baseline level of trust predisposes a person's initial response to an automated system, independent of specific interactions. *Situational* trust, on the other hand, pertains to trust levels influenced by immediate external factors, including the complexity of the task, environmental conditions, and situational demands. These elements affect the degree to which trust influences reliance on an automated system, with considerations such as workload and perceived risks or benefits of using the system playing crucial roles. *Learned* trust pertains to trust that develops from personal experiences with a particular automated system. This form of trust is dynamic and history-dependent, shaped by the user's direct interactions with and evaluations of the system's performance. Hoff and Bashir (2015) further categorize learned trust into initial and dynamic segments. *Initial learned* trust is based on pre-existing knowledge and perceptions before interaction, influenced by factors such as past encounters with similar technologies or the system's reputation. *Dynamic learned* trust evolves from continuous use, fluctuating in response to the system's performance over time.

Hoff and Bashir's (2015) model represent an important synthesis in the field of trust in automation, their work mainly outlines an overview of the factors at each layer. Moreover, this model, much like Lee and See's (2004) work, has yet to be extensively validated through empirical research.

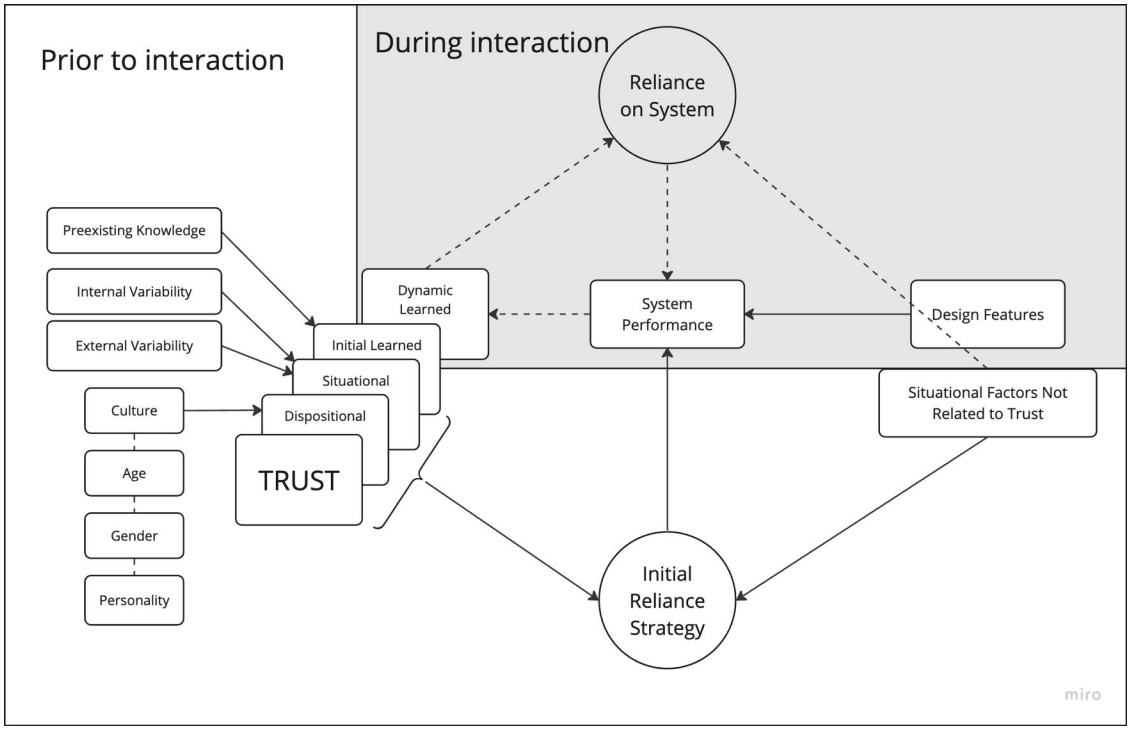


Figure 9, Three-layered model of Trust in Automation (Hoff and Bashir, 2015)

In addition to the aforementioned conceptual models, several computational and mathematical models have been developed to offer a quantitative account of trust, some of which are described here. Gao and Lee (2006) developed the Extended Decision Field Theory (EDFT), a dynamic-cognitive framework aimed at explaining the evolution of preferences within decision-making contexts under uncertainty. Utilizing an autoregressive model, EDFT integrates past preferences and new information to estimate current preference shifts, mapping the dynamics of trust and self-confidence in automation contexts. This model applies a segmented function to formulate beliefs about automation capabilities versus manual control, aligning with empirical observations on trust inertia and the interplay between trust, self-confidence, and reliance on automation. Xu and Dudek (2015) explored the relationship between trust and reliance through the Online Probabilistic Trust Inference Model (OPTIMO), a Dynamic Bayesian Network designed to quantify an individual's trust level in robotic teammates, suggesting reliance as a tangible measure of trust. Akash et al. (2017) proposed a linear model defining trust evolution, which accounts for human biases influenced by past experiences and expectations. This model quantifies trust adjustments based on the discrepancies between current trust and experiences, cumulative trust, and expected biases, addressing the challenge of applying these principles broadly due to its reliance on direct trust inquiries from users. Yang et al. (2017) applied a first-order linear time-invariant system to analyze how average trust in automation reaches equilibrium over repeated interactions. Following this, Guo and Yang (2021) introduced a personalized trust prediction model using

Bayesian inference and a Beta distribution to assess trust dynamics, acknowledging the model's assumption on static automation capabilities and its impact-based dichotomy of automation performance. Furthermore, Lewis and Weigert (2012) emphasize the significance of historical feedback loops in developing trust relationships. Jonker and Treur (1999) further investigate trust dynamics, emphasizing the reciprocal evolution of trust based on interactions. Manzey et al. (2012) identified positive and negative feedback mechanisms influencing trust adjustments, noting the disproportionately larger impact of negative experiences on trust recalibration compared to positive interactions, as reinforced by subsequent studies (Yang et al., 2016).

Up to this point, the dissertation has endeavored to deliver a detailed overview of the research background and context. This includes an in-depth exploration of the theoretical constructs central to the study, such as Supervisory Control, Levels of Automation (LOA), Trust in Automation (TiA), and various TiA models. The extensive range of topics, concepts, and theoretical frameworks underscore the inherent complexity of intertwining diverse subjects and constructs. Significantly, given this dissertation's emphasis on the modeling facets of Human-Automation Interaction (HAI), it becomes crucial to thoroughly comprehend the field and disciplinary context within which this research is embedded. An integral part of this exploration involves dissecting the nature of complexity and examining how models can effectively navigate and address such complexity in the realm of HFE. The forthcoming chapter will delve into the theoretical foundations of the dissertation, concentrating on the broad discipline of Human Factors and Ergonomics (HFE). It will explore the historical evolution and defining characteristics of HFE. This foundational knowledge will inform the subsequent evaluation of existing models and modeling efforts within the field, as well as assess the appropriateness of various modeling approaches in meeting the specific needs of the HFE discipline.

3 Theoretical Grounding

3.1 Evolution of Human Factors and Ergonomics (HFE)

The discipline of Human Factors and Ergonomics (HFE) originated from the old idea of ‘fitness for purpose' and progressed particularly in response to the design and operational challenges posed by technological advancements in the 20th century (Green & Jordan, 1999; Meister, 2000). Throughout its history, the discipline has been known by various monikers, such as human engineering and engineering psychology, though it is most commonly referred to as 'human factors' in the United States and 'ergonomics' in the United Kingdom and Europe (Dempsey et al., 2000; Wickens et al., 2004). The term 'human factors' became prevalent in the late 1940s, when the field's development was significantly influenced by psychology and engineering (Meister & Enderwick, 2001). The emphasis was placed on human-centric design and incorporating human considerations throughout the system design process (Sanders & McCormick, 1998).

With tremendous progress during and post-World War II, Human Factors and Ergonomics (HFE) began to flourish globally, finding practical applications across various industrial contexts (Waterson & Eason, 2009). In the 1960s, the discipline secured a solid footing in both academic circles and the industrial sector, progressing towards a more integrated relationship with stakeholders in the civil, governmental, and industrial realms, including users and practitioners. Internationally, HFE increasingly played a central role in shaping health and safety regulations (Moray, 2008). This period marked a significant expansion in HFE, characterized by advancements in consumer ergonomics, the establishment of standards, the integration of automation and systems ergonomics, the rise of computing and technological innovations, and the refinement of job and work design methodologies (Moray, 2008; Waterson & Eason, 2009). The scope of consumer ergonomics broadened beyond seating solutions to encompass a diverse array of consumer goods, ranging from household devices to kitchen layouts and the design of hospital beds (Waterson & Eason, 2009).

During the 1960s, the contributions of Human Factors and Ergonomics (HFE) to computer systems primarily revolved around the design of interface hardware, such as keyboards. It was not until the advent of the personal computer (PC) in the post-1970 era that comprehensive empirical studies began to focus on computer software (Meister, 1999). The rise of interactive computing brought to the forefront a multitude of human-centric issues, prompting ergonomists to actively engage in research, evaluation, and design. Initially, interactions with computers were grounded in programming languages tailored for computer specialists, proving unsuitable for a broader audience including accountants, clerks, engineers, and managers. This necessitated the creation of software interfaces designed to be intuitive and user-friendly for the growing demographic of PC users. The introduction of personal computers heralded the era of ‘point and click’ graphical user interfaces, setting a new benchmark for human-computer interaction (Waterson & Eason, 2009).

The introduction of the computer revolution in the 1980s thrust Human Factors and Ergonomics (HFE) into widespread attention, highlighting the importance of ergonomically designed computer hardware, user-centric software, the role of human factors in office environments, and the broader implications of technological advancements on individuals (Sanders &

McCormick, 1993). A pivotal focus of research within this period was the operation of nuclear power plants, a direction largely motivated by the notorious nuclear incidents at the Three Mile Island facility in the US in 1979 and the Chernobyl facility in 1986. These incidents, along with other technological catastrophes such as the Bhopal disaster in India in 1984 and the Phillips Petroleum plant explosions in Texas in 1989, accentuated the profound consequences of neglecting human factors, both in human casualties and financial ramifications. Meshkati (1991) conducted analyses on these tragedies, identifying a common thread of insufficient human factors considerations contributing to the magnitude of these disasters. In response, there was an intensive effort towards enhancing facility safety, notably through the integration of HFE programs aimed at augmenting operator support and ensuring the prevention of similar disasters (Meister, 1999).

Over recent decades, the Human Factors and Ergonomics (HFE) field has witnessed exponential growth in its literature, including books, scholarly articles, and conference proceedings (Meister, 1999). Karwowski (2012) observed that HFE has broadened its horizons, extending beyond the traditional domains of physical, psycho-physiological, cognitive, and organizational/macro ergonomics to embrace systems-oriented approaches globally. This evolution mirrors the growing intricacies of human-system interactions. Notably, the discipline has ventured into emerging areas such as nanoergonomics and neuroergonomics, a shift towards the human-centered design of increasingly complex systems (Karwowski, 2005; Parasuraman & Rizzo, 2008). Research in cognitive ergonomics, human-computer interaction, organizational design and management, and the relationship between work and health has seen substantial growth (Waterson, 2011). The scope of Human Factors and Ergonomics (HFE) has widened, embracing new realms such as the effects of information and communication technology on work and daily activities (Dul et al., 2012), interventions addressing psychosocial risks in workplaces (Petit et al., 2011), and fostering sustainability in energy, waste management, and transportation (Haslam & Waterson, 2013).

3.2 HFE As Scientific Discipline

Meister (1999) argues that Human Factors and Ergonomics (HFE) is grounded in the pursuit of generalization and prediction, essential traits of scientific inquiry. Chapanis (1988) delves deeper, articulating that generalizability is about extending research outcomes beyond the original conditions of study. This ability to generalize from observed events is critical for preparing and managing unencountered human behaviors or scenarios. Sanders and McCormick (1998) emphasize that HFE is anchored in the scientific method, leveraging objective data to evaluate hypotheses and gather fundamental insights into human behavior. Research within HFE aims to uncover and understand the psychological, social, physical, and biological facets of humans, with the goal of integrating this knowledge into the design and operation of products or systems. This integration seeks to enhance human efficiency, well-being, safety, and comfort in their interactions with various environments and technologies (Stramler, 1992). In HFE, the study of human behavior extends across various environments, from highly controlled laboratory settings to real-world systems, to achieve refined observations and experimental results. These environments are chosen to reflect the complexity of systems relevant to the research, ranging from controlled labs facilitating precise

observations to naturalistic settings where normal behavior, incidents, and accidents of users can be observed (Wickens et al., 2004). According to Wickens et al. (2004), a thorough understanding, alongside the ability to generalize and predict human behavior, is greatly enhanced by employing a blend of diverse observational methods and analytical techniques.

Contrary to viewing Human Factors and Ergonomics (HFE) strictly as a scientific discipline, some scholars advocate for its recognition as an art or craft, employing scientific methods as a means to an end rather than the end itself. Moray (2000, p. 529) emphasized the importance of understanding HFE within the complex social contexts where human work takes place, stating "*our discipline is an art not a basic science, and one which only makes sense in the full richness of the social setting in which people work*." He highlights that the application of HFE encompasses a wide array of contextual factors including team dynamics, individual motivations, organizational culture, and the varied purposes and scales of systems under consideration. Wilson (2000) illustrates the multidimensional nature of HFE in the domain of human-computer interaction, acknowledging it as a confluence of craft, science, and engineering. This perspective explains the craft in HFE's aims to implement and evaluate, its scientific aspect in explaining and predicting behaviors, and its engineering facet in designing systems for enhanced performance. Such viewpoints advocate for a universal approach to HFE, recognizing it as a discipline that transcends traditional boundaries by integrating artistic craftsmanship, rigorous scientific inquiry, and pragmatic engineering solutions.

Despite the diverse perspectives on the nature of Human Factors and Ergonomics (HFE), there is a prevailing consensus among scholars that it fundamentally qualifies as a science. Human Factors and Ergonomics (HFE) as a scientific discipline focuses on understanding interactions between humans and various elements in the environment. It is also a professional field that applies theoretical principles, data, and methods to design, with the aim of enhancing well-being and overall performance (Dul et al., 2012). The primary goals of HFE are to improve the efficiency and effectiveness of work and other activities while promoting key human values such as safety, reduced fatigue and stress, and an enhanced quality of life (Sanders and McCormick, 1993). To realize these objectives, many experts have emphasized the importance of knowledge transfer and the creation of synergy between HFE research and its practical application (Caple, 2008; Meister, 2000; Sind-Prunier, 1996; Singleton, 1994). This involves bridging the gap between theoretical research and practical implementation, ensuring that insights from research are effectively translated into tangible improvements in design and practice. This collaboration and integration of theory and practice are essential for the advancement of HFE, making it a useful and impactful discipline. Several experts in Human Factors and Ergonomics (HFE) have underscored the vital role of theory in the discipline. Meister (1999), for instance, viewed the connection between theory, research, and practice into three distinct segments: (1) The relationship between theory and research; the idea that theoretical frameworks provide a foundation upon which research is built and directed. (2) The relationship between research and practice, which focuses on the idea that research should offer practical guidelines for design and operation. This implies that the findings and insights gained from research should directly influence and shape practical applications in the field of ergonomics. (3) The interrelationship among theory, research, and practice, implying that these three elements are interdependent. The absence of a solid theoretical base would mean that research cannot effectively provide the guidelines necessary to inform practice. Expanding on

the importance of theory in HFE, Getty (1995) emphasized the necessity for ergonomics principles to be grounded in robust and validated research. This approach is crucial for the scientific integrity and long-term development of the discipline. Karwowski (2005) took this notion a step further by outlining three primary paradigms within HFE: (1) Ergonomics theory, which involves understanding, describing, and evaluating interactions between humans and systems. (2) Ergonomics abstraction, that utilizes insights about human-system interactions to make testable predictions against real-world scenarios. (3) Ergonomics design, that focuses on applying knowledge of human-system interactions to create systems that not only meet consumer needs but also adhere to human compatibility requirements.

3.3 HFE As Basic and Applied Discipline

Basic science is driven by the quest to answer fundamental questions out of pure interest, aiming to unravel the underlying mechanisms of various processes without any commercial intent (Horrobin, 1969; Rimnac & Leopold, 2014). It often begins with unique observations and a genuine pursuit of knowledge (Nudds & Villard, 2006). Historical examples include Michael Faraday's development of electromagnetic induction principles in 1821 and Heinrich Hertz's discovery of what is now known as radio waves in 1886. Such research, while foundational, seldom transitions directly into practical applications, as its commercial potential is not immediately evident. On occasions where basic research does prove to be of practical use, its applications are frequently realized in fields far from the original study (Horrobin, 1969). In contrast, applied science leverages the insights gained from basic science to push technological, material, or treatment advancements forward (Rimnac & Leopold, 2014). This approach is often motivated by the need to address pressing industry problems (Nudds & Villard, 2006), creating a bridge between theoretical knowledge and real-world applications.

Human Factors and Ergonomics (HFE) emerged as a distinct field through the collaborative efforts of applied scientists addressing multifaceted challenges that spanned various disciplines (Bridger, 2009). Recognized for integrating principles from anatomy, physiology, and psychology, HFE has also established significant ties with practical fields such as medicine and engineering (Singleton, 1994; Wilson, 2000). Initially focusing on human interactions with physical devices in military and industrial contexts, the domain of HFE has significantly expanded, reflecting a shift towards a more holistic consideration of human interactions across a broader range of environments and systems. This evolution indicates the inherently goal-driven nature of HFE, which, unlike the predominantly technical focus of engineering, prioritizes the human element in the design process, emphasizing the impact of design on human interactions with products, environments, and systems (Sanders & McCormick, 1998).

At its core, HFE is fundamentally linked to design, whether it pertains to work practices, products, or entire systems (Green & Jordan, 1999). The discipline's foundation lies in the strategic application of knowledge about human traits to foster compatibility within interactive systems involving people, machines, and their environments (Karwowski, 2012). The aim is to adapt the design of tools, environments, and systems to better suit human needs, capabilities, and limitations (Sanders & McCormick, 1998). Originally, HFE addressed singular issues of individual interactions with machines or specific environmental factors. However, the complexity of contemporary life demands a more sophisticated understanding of human

behavior and performance, extending beyond the constraints of traditional ergonomics (Wilson, 2000).

Despite its practical orientation, Wilson (2000) argues for the necessity of both basic and applied research within HFE to facilitate evidence-based practice and ensure the discipline's contributions are both meaningful and empirically grounded. This approach also becomes critical in integrating research findings into practical applications to enhance the design and usability of systems and environments for human use. Turner (2002) posits that an ideal research framework should both guide practice and foster theoretical advancements, with practice, in turn, benefiting from research insights and generating further inquiries for exploration. The synergy between research and practice is envisioned to establish a robust theoretical foundation, fostering the growth and development of a professional community (Haddow & Klobas, 2004). Despite these aspirations, the practical integration of research findings into professional practice frequently fails to meet these expectations, revealing a gap between theoretical ideals and operational realities.

3.4 HFE As System Discipline

Dul et al. (2012) assert that HFE primarily focuses on systems where humans interact with their environment. Yet, the term 'system', while commonly used, often lacks a clear, explicit definition, leading to varied interpretations and applications. Simply classifying a discipline as "systems-based" does not enhance understanding or provide clearer descriptions. According to Dul et al. (2012), in HFE, a system is defined as a set of interrelated independent parts or elements, with the acknowledgment that the whole is more than just the sum of its parts. Dynamic systems, on the other hand, change their system *state* with time. These states can be number of students enrolled in a class, population of a country, physical and mental activities, or psychological constructs such as one's trust. Singleton (1974) suggested that dynamic systems consist of interconnected objects that evolve over time, and for human-made systems, they serve a specific purpose. While this definition might be debated in other sciences, particularly regarding natural systems and their purpose, it provides a foundation for understanding systems. Chapanis (1996) defined HFE system as an interactive combination of people, materials, tools, machines, software, facilities, and procedures, all working together for a common purpose. Wilson (2014) proposed that a system is a set of interconnected activities or entities (including hardware, software, buildings, spaces, communities, and people) with a shared purpose. These entities are linked through various forms of state, function, and causality, and the system evolves in response to different circumstances and events. A system is conceptualized as existing within a boundary, having inputs and outputs with potential many-to-many connections. In accordance with Gestalt principles, the system as a whole is typically more significant (useful, powerful, functional, etc.) than the mere sum of its parts (Wertheimer, 2012).

Systems can be categorized based on different features, however, the focus in this dissertation is on two distinct features of systems, complex and dynamic systems. Bossel (2007) contends that inherently, all systems possess dynamic qualities, including those that seem relatively static at first glance. Nonetheless, the designation "Dynamic System" is specifically allocated for

systems that undergo state changes and, as a result, exhibit dynamic behavior as time progresses. The terms complex and complexity require more clarity, as discussed in the next section.

3.4.1 Complicated vs. Complex Systems

The introduction of a social element, such as human interaction, transforms a system from merely complicated to complex (Cilliers, 2008; Dekker et al., 2011). In complex systems, the parts cannot be understood in isolation from the whole. According to Ottino (2003), the system itself should be the primary focus of analysis. This principle emphasizes the importance of understanding the interconnectedness and emergent properties of systems rather than merely analyzing their individual components in isolation.

It seems that the distinction between "complex" and "complicated" systems is necessary. Complicated systems, such as container ships, are intricate and composed of many parts. Despite their intricacy, they can be disassembled and reassembled, and are, in principle, understandable and describable, even if not by a single individual. This characteristic categorizes them as complicated. In contrast, complex systems are defined by the interactions among their components. A container ship transforms into a complex system when integrated into the real world, encompassing factors such as cultural diversity, communication styles, varying bridge hierarchies (Orasanu & Martin, 1998), fatigue effects, procedural implications (Snook & Irvine, 1969), diverse training and language standards (Hutchins, 1996), and cross-cultural differences in risk perception and behavior (Lund & Rundmo, 2009). These elements transcend engineering specifications and reliability predictions, making the system complex.

Complexity in behavior stems from the interplay among system components, emphasizing the importance of their interrelations rather than the components themselves (Dekker et al., 2011). System properties emerge from these interactions, not residing in any single component (Israel, 2005). Complex systems can develop new structures internally, independent of external design. They adapt their internal structures in response to environmental changes (Urry, 2006). Complexity is an attribute of the system as a whole, not of its components (Zuchowski, 2018). Each component's understanding is limited and localized, with no single component capable of encapsulating the system's entire complexity. Thus, the system's behavior cannot be simplified to the behaviors of its components.
In complex systems, connections are local, and each component operates without full knowledge of the system's overall behavior. Components react based on their immediate information, leading to complexity from the vast network of interactions and relationships stemming from these local responses. With interdependencies and interactions expanding rapidly, the complex system's boundaries become blurred (Mitchell, 2006). Furthermore, complex systems are influenced by their history and path dependence, which extends beyond their boundaries. Their current behavior is shaped not only by their own past but also by the history of surrounding events (Sterman, 1994). In complex systems, conditions are irreversible, meaning it is impossible to completely reconstruct the specific circumstances that led to a particular outcome, such as an accident (Dekker et al., 2011). These systems are in continuous flux, with evolving relationships and adaptations to their environment. As a result, the state of

a complex system post-accident is never the same as it was before the accident; it changes not only due to the accident but also due to the passage of time. This constant change limits the predictive accuracy of retrospective failure analyses (Leveson, 2002). In addition, the process of reconstructing events in a complex system post-event is filled with challenges, not only because of the nature of complexity but also due to psychological factors such as hindsight bias that can distort past events (Fischhoff & Beyth, 1975; Hugh & Dekker, 2009). The system under scrutiny after an event is never the same system that produced the outcome.

Closely related to the concepts of complexity and systems theory is Nonlinear Dynamic Systems (NSD) theory (Guastello, 2017) which captures the essence of complex dynamic systems and their fundamental properties.

3.4.2 Nonlinear Dynamic Systems (NDS)

Nonlinear Dynamic Systems (NDS) theory functions as a general systems theory due to its broad applicability across a diverse array of phenomena. NDS has effectively incorporated and expanded upon several key concepts from general systems theory. Among these are the notions of feedback loops and mathematical formalisms, which form the core elements of NDS applications (Guastello & Liebovitch, 2009). Feedback loops, fundamental to understanding system behaviors and interactions, are especially crucial in NDS as they help explain how systems self-regulate and evolve over time. Mathematical formalisms provide the necessary structure to model and analyze the dynamic behavior of systems in a precise and quantifiable manner. Nonlinear Dynamic Systems (NDS) theory is grounded in four fundamental principles that redefine traditional notions of system analysis and behavior:

1. *Variability and Deterministic Functions*

NDS posits that seemingly random events can be produced by simple deterministic functions, though identifying these functions can be challenging. The focus in NDS is as much on the analysis of variability as it is on the analysis of the underlying structure of variability, focusing on understanding the size and structure of variability, the processes generating dynamic patterns, and identifying system variables that influence these outcomes (Fuchs, 2013).

2. *Diverse Types of System Changes*

Rather than solely using assumptions of linearity, NDS acknowledges a variety of change types. These are represented through multiple modeling structures (Guastello & Liebovitch, 2009; Sprott, 2003). Temporal patterns in NDS are seen as indicators of specific underlying dynamics, where time typically acts as an implicit variable. The state of an agent at one point in time is a nonlinear function of its state at a previous time and other influencing control variables.

3. *Dynamics of Stability and Instability*

Contrary to the belief that systems exist in equilibrium until disrupted by external forces, NDS suggests that systems inherently produce stabilities, instabilities, and other change dynamics as part of their internal functioning (Jost, 2005; Thelen, 2005). Both internal and external disturbances can influence a system, and the focus is on identifying variables that govern the system's reactivity to these disturbances.

4. *Control and Emergence Over Traditional Causality*

In NDS, the concept of causality is replaced with ideas of control and emergence. Instead of linear cause-effect relationships (where Event A causes Event B), NDS views each agent as following a dynamic path influenced by its prior state. The impact of variables on this path depends on the agent's or system's previous state, leading to outcomes where small influences can have significant effects and vice versa (De Bot, 2017). Emergent phenomena in NDS are seen as outcomes of complex system behaviors and interactions, involving both bottom-up and top-down processes. These emergent phenomena require viewing situations through an appropriate lens to be understood correctly (Guastello, 2017).

With that being said, Human-automation interaction (HAI) exemplifies a nonlinear dynamic and complex system in two distinct ways (Fereidunian et al., 2015; Karwowski, 2012). First, the tasks jointly executed by humans and automated systems are fundamentally complex and require adaptation to evolving environmental conditions. Second, the interaction between humans and automation systems is, in itself, a multifaceted phenomenon that requires continuous adaptation to changes in the system's states. HAI depends on and reacts to initial conditions; thus it cannot be rule-governed and must be a case of continuous adaptation (Flach, 2012; Hollnagel, 2021).

The increasing complexity of sociotechnical systems necessitates a paradigm shift in the scientific development and practice of Human Factors and Ergonomics (HFE), a sentiment echoed across various disciplines and scholars (e.g., Allen & Varga, 2007; Dore & Rosser Jr, 2007; Fleener & Merritt, 2007; Guastello, 2007; Karwowski, 2012; Walker et al., 2017; Zausner, 2007). This shift is not just a mere change in techniques or tools; it represents a fundamental transformation in how phenomena are understood, approached, modeled, and explained. A new scientific paradigm in HFE would entail the emergence of fresh concepts for interpreting complex phenomena, posing novel questions, and developing new methodologies aligned with these questions. This paradigm shift would also bring about new explanations for phenomena, offering insights that might have been previously fragmented or not fully understood. Moreover, it would significantly alter perspectives, leading to an enhanced understanding of existing phenomena.

It is important to recognize that all research techniques and methods have their strengths, opportunities, and limitations (Frankfort-Nachmias et al., 2014). This diversity in methodological approaches means that some techniques are more suited to certain types of problems than others (Hughes & Sharrock, 2017). Therefore, the selection of appropriate methods and techniques is crucial for effectively understanding and solving the complex problems faced in modern ergonomics.

4 Research Methods

This chapter presents the overall research methodology adopted throughout this doctoral dissertation. It primarily concentrates on the role of models and modeling complex systems, as well as simulation techniques employed in this dissertation. This section also argues the justification for selecting simulation as an effective modeling strategy to tackle the dynamic complexities inherent in Human-Automation Interaction (HAI) research.

4.1 Theory, Model, and Modeling

Theory is viewed as a structured, explanatory, abstract, and coherent framework comprising interconnected statements about a certain aspect of reality. These statements consist of constructs linked together through testable propositions, underpinned by a logical structure and certain assumptions (Davis et al., 2007; Dubin, 1970; Fawcett, 1988). The constituent elements of theories often consist of well-formalized models. Theories that arise from a process of theorizing rooted in an explicit, formal model have the potential to be more robust and far-reaching compared to theories primarily based on implicit mental models (Wacker, 2004).
Methodologically, this Doctoral research focus lies in creating a formalized, structural model, utilizing computer-assisted simulations. This approach is also referred to as computer-supported theory-building (Hanneman, 1988). However, not every model qualifies as a theory. To qualify as a theory, a model must be accompanied by a plausible explanation of why it generates the observed behavior (Lane & Schwaninger, 2008).

Models can be represented through various forms such as verbal descriptions, graphical illustrations (e.g., diagrams, and images), mathematical formulations, physical representations, or a combination thereof, such as computational models that utilize mathematical equations to generate numerical data and graphical outputs (Sheridan, 2017). Models can also be categorized in various ways based on different criteria, including dynamic versus static, deterministic versus stochastic, simple versus complex, and so on. A critical distinction, especially in terms of model validity, lies between "causal-descriptive" (explanatory, theory-like, or "white-box") models and purely descriptive or "correlational" (data-driven or "black-box") models (Barlas, 1996). Purely correlational, or black-box, models do not assert causal structures. In these models, the primary focus is on the aggregate output behavior. A model of this type is deemed valid if its output aligns with real-world output within a certain accuracy range, without delving into the internal validity of the model's individual components or relationships. This form of "output" validation often resembles a classical statistical testing problem. Data-driven models are primarily used for forecasting, such as time-series or regression models. Conversely, explanatory, causal-descriptive, or white-box, models are assertions about the actual processes of real systems. For these models, producing an accurate output is not the only criterion for validity. The internal structure's validity is paramount. As these models serve as theories about the real system, they must not only replicate or predict the system's behavior but also explain how this behavior arises and possibly suggest modifications to alter the existing behavior (Barlas & Carpenter, 1990).

That said, the formalization of any model involves a crucial step: specifying the nature of the relationships among the components within the state space. The primary goal of developing formal models of dynamic processes is to comprehend the dynamic outcomes arising from the interactions among the components within the state space. In models characterized by extensive state spaces, marked by strong interconnections, nonlinearity, time delays, noise, and feedback mechanisms among the states, deducing the dynamic behavior of the theory can be exceedingly difficult. In such complex scenarios, the most effective means of understanding and analyzing the theory is through simulation modeling (Hanneman, 1988; 1991). The primary value of a simulation outcome is that it presents propositions that can be tested and, if necessary, refuted. The focus is not just on whether a proposition is true or false but on providing a foundation for constructing arguments.

4.2 Simulation Modeling

Simulation is essentially the practical application or execution of a model, typically through numerical methods or computation. Its main objective is to simplify the complexity of the analytical model, converting the theory into a manageable form, often via computerization (Beisbart, 2012). Modeling and computer simulation of dynamic systems have the advantage unachievable from only initial knowledge about the system. Complex behaviors such as oscillations, collapses, or chaotic patterns, which are not directly inferable from the individual elements and their interactions, become apparent through simulations. Predicting the response of dynamic systems under specific conditions can be very challenging. In these situations, the goal is to generate reliable insights into the system's behavior (Bossel, 2007). A computer model capable of simulating system behavior provides insights into how the system might react under different circumstances. Through comprehending these behaviors, one can establish various scenarios and consider the appropriate interventions needed (Lane & Schwaninger, 2008). Additionally, simulations are particularly valuable for scientific hypothesis testing. Formulating competing hypotheses into program statements, integrating them into a simulation model, and comparing the outcomes with actual observations can be a straightforward and effective method (Schwaninger & Grösser, 2008).

Simulation modeling, since its large-scale inception in the 1940s, has become a fundamental tool across various disciplines. Simulation is particularly useful for analyzing systems where it is impractical to systematically check every possible state of a model (Bratley et al., 2011). Although the application of simulation modeling is widespread, each field has developed its own specialized techniques, tools, and terminology. For instance, computer scientists often employ discrete-event models, which concentrate on changes in state (Robinson, 2005). This approach is also prevalent in human performance models (Choi, 2018), where the focus is on discrete events that alter the system's state. In contrast, agent-based models prioritize the decision-making processes of independent actors. These models are particularly effective for simulating the behaviors of human groups, capturing the dynamics of individual and collective actions (Pan et al., 2007). Engineers, on the other hand, might opt for continuous-event models to represent systems with states that change constantly (Robinson, 2005).

Dynamic systems and their corresponding modeling approaches can be characterized by several contrasting properties, essential for selecting the most appropriate modeling strategy (Bossel, 2007). Some of the key properties are briefly discussed here:

1. *Explanatory vs. Descriptive:* Simulating behavior can be approached in two ways. The first approach involves deriving a behavioral *description* based on observations of one or several similar systems, noting how they respond under varying conditions, and then applying mathematical relationships to correlate inputs with outputs, thus replicating the real system's behavior. The second method entails attempting to *explain* behavior through modeling the actual processes within the real system, which requires extensive knowledge about the structure of the system itself.
2. *Real Parameter vs. Parameter Fitting:* Explanatory models, aiming to reflect the real system's structure closely, employ actual system parameters, which might be directly measurable or derivable from existing studies. When direct measurement is infeasible, parameter fitting becomes essential, aligning the model's quantitative outputs with observed system behaviors. However, parameter fitting in explanatory models, which maintain structural validity, is preferable to fitting in descriptive models, which might incorporate structurally inaccurate relationships.
3. *Deterministic vs. Stochastic:* Deterministic models exclude random parameter changes, interactions between system elements, and environmental influences, portraying a predictable, unvarying system behavior. In contrast, stochastic models explicitly incorporate randomness, for instance, through transition probabilities or environmental variability, leading to divergent outcomes with each simulation run. Monte Carlo simulations and other stochastic methods provide insights into behavior distributions, mean values, and variance. When models aim to represent the collective behavior of numerous entities or processes, individual stochastic behaviors and processes are averaged to depict a consolidated outcome. This approach transforms the unpredictable nature of individual elements into a more predictable aggregated behavior, often permitting the use of deterministic models to approximate real system dynamics effectively.
4. *Constant Parameters vs. Time-Variant Parameters:* System parameters can be constants or time-dependent functions. In systems with constant parameters, the structure and interrelations remain unchanged over time, leading to repeatable outcomes under identical conditions. Conversely, systems characterized by time-variant parameters evolve, reflecting changes in structure, relationships, and consequently, behavior over time.

Understanding these distinctions and properties is crucial for selecting and developing a modeling approach that accurately represents the dynamics of the system under study while meeting the research or application objectives.

Structural models, due to their emphasis on the foundational relationships within a system, offer a powerful tool for simulating responses to novel conditions, exploring a wide spectrum of potential behaviors and developmental trajectories, and understanding the conditions and

opportunities for systemic change (Größler et al., 2008). While there is a theoretical distinction between descriptive and explanatory models, in practice, purely explanatory models are rare. They often rely on descriptive sub-models to aggregate and delineate individual relationships (Bossel, 2007).

This dissertation adopts two distinct modeling approaches: Fuzzy Logic and System Dynamics simulations, each tailored to address specific research challenges and phenomena under investigation; that is, Levels of Automation (LOA) and Trust in Automation (TiA). The selection of these methodologies reflects their unique advantages, foundational assumptions, and developmental frameworks, which are elaborated upon in the subsequent sections.
In the context of model evaluation and validation, this doctoral dissertation delves rather deeply into the philosophical underpinnings of scientific inquiry, resulting in the formulation of a set of evaluative criteria detailed in the subsequent chapters. Given the critical importance of this theme – a cornerstone undergirding the entire dissertation – a dedicated chapter has been allocated to thoroughly explore the philosophy of science and the principles underpinning model validity. This focused approach is necessitated by the direct relationship between the philosophical and epistemological foundations with the development and assessment of models, as critically examined in Articles 1 and 2 of this dissertation. These sections aim to provide an extensive examination of the theoretical aspects, ensuring a comprehensive understanding of the processes involved in model evaluation and the significance of their philosophical dimensions.

4.2.1 Fuzzy Logic (FL)

At the heart of most human communication is an element of fuzziness—statements are often imprecise, lacking in clarity, and not perfectly defined. Fuzzy logic mirrors this aspect by recognizing that natural language terms can apply with varying degrees of relevance across different objects or situations. This characteristic makes fuzzy set theory an effective tool for modeling the way humans approach decision-making and inference.

Originating from Zadeh's (1965) seminal work, fuzzy logic extends the classical set theory to accommodate ambiguity and subjectivity, enabling the mathematical representation of vague relationships through fuzzy sets and membership functions. This approach facilitates the expression of partial membership, where the degree to which an entity belongs to a set, ranges between zero (no membership) and one (full membership), effectively capturing the concepts that do not fit neatly into binary categories. Contrasting with classical set theory where a temperature of 15°C might strictly be classified as cold, fuzzy logic allows for a more comprehensive interpretation—deeming 15°C as simultaneously 65% cold and 35% warm, which shows the subjective and variable nature of such categorizations. The selection of an appropriate membership function (e.g., triangular, trapezoidal, or Gaussian curves) is key in the construction of a fuzzy logic system, influencing its effectiveness and simplicity (Ross, 2005).

At the core of Fuzzy Logic are the Fuzzy rule-based inference systems that establish connections between input and output variables using if-then logic, where the inputs are assessed across a continuum and associated with degrees of membership in relevant fuzzy sets.

These systems determine the extent to which a rule applies based on the aggregated membership degrees of input values, with the resultant fuzzy output shaped by these degrees through operations such as truncation or scaling. Utilizing fuzzy linguistic evaluations, membership functions, fuzzy operators, and if-then statements, Fuzzy Inference Systems (FIS) offer a structured framework for reasoning under uncertainty (Dzitac et al., 2017). Predominantly, there are two types of FIS: Mamdani and Sugeno, each outlining a distinct method for generating outputs (Ross, 2005). In this dissertation, the Mamdani approach is employed. The Mamdani method operates by transforming the "or" and "and" connectives in rules to "max" and "min" operators, respectively, ensuring that the combined outputs follow a coherent aggregation logic. This approach facilitates the generation of a composite output from multiple fuzzy membership functions, presented across the output variable's universe of discourse. The process of converting a precise input vector into a corresponding output vector via fuzzy rules unfolds through a three-stage process, Fuzzification, Fuzzy Inference, and Defuzzification.

Fuzzification initiates the fuzzy inference process by translating crisp input value into corresponding fuzzy linguistic variables, establishing a framework where every input component is matched with a linguistic variable and its set of linguistic values defined across the input's universe of discourse (Ross, 2005). Membership functions are allocated to each linguistic value, enabling the conversion of crisp inputs into a spectrum of membership values, which effectively bridges the gap between crisp data and fuzzy logic interpretation (Ross, 2004). The essence of a fuzzy system unfolds in the rule establishment stage, where knowledge about the subject matter is encapsulated in conditional if-then statements. These rules, which form the core of the Fuzzy Inference System (FIS), articulate the implications of specific conditions (antecedents) leading to certain outcomes (consequents). The evaluation of these conditions is quantified through the minimum or maximum aggregation of the membership values, determining the strength of the rule's application (Van Leekwijck & Kerre, 1999). The derivation of conclusions in a fuzzy inference system involves integrating the outcomes of multiple rules, each contributing to the overall output based on the degree of fulfillment of their antecedents. This aggregation process concludes in a composite output fuzzy set for each output variable, encapsulating the cumulative wisdom of the system's rules. Defuzzification is the final transformative step, where the fuzzy output is distilled into crisp, actionable values. Various methods exist for this purpose, with the Height method selected in this dissertation for its simplicity and straightforward implementation. This process calculates the weighted average of the maximum membership values to derive a precise output value for each output variable.

Fuzzy logic offers a practical contribution to handling ambiguous information, particularly in modeling systems where precise definitions are rather difficult to articulate. This approach thrives on embracing the indeterminacy and subjectivity inherent in many decision-making processes (Kahraman et al., 2006). Fuzzy set theory provides a robust framework for the quantitative handling of imprecise problems encountered in the existing levels of automation. Its core strength lies in treating vague parameters (e.g., information acquisition, information analysis, decision selection, action implementation) as fuzzy values rather than attempting to force them into a precise framework, thereby yielding more reliable outcomes (Konstandinidou et al., 2006).

4.2.2 System Dynamics (SD)

Jay W. Forrester, from MIT's Sloan School of Management, proposed the concept of "system dynamics" in a note to the Faculty Research Seminar in 1956, marking a significant departure from traditional economic modeling (Forrester, 1987, 1997). He criticized existing economic models for not capturing the essential feedback loop structures, neglecting the integrated flows of goods, money, information, and labor, and failing to account for the changing mental models impacting economic processes. Forrester pointed out the limitations of linear equations, the constraints of model building due to computational capacities, the overreliance on multiple regression for deriving economic behavior coefficients, and the lack of critical reflection on underlying model assumptions (Forrester, 1997). In contrast, Forrester advocated for the utilization of servomechanisms, differential equations, and simulation techniques. He envisioned models with dynamic structures to closely examine system actions and their driving forces, emphasizing the significance of lags and time delays. Forrester delineated system dynamics through a four-layered hierarchy: feedback loops forming the system's structural basis, stock variables indicating accumulations within these loops, flow variables depicting activities, and a framework for aligning system goals with observed states to inform actions (Forrester, 1992, 2012).

Endogeneity is a hallmark of system dynamics, positing that within-system causal influences, encapsulated in feedback loops, drive system behaviors—contrasting with models where causes are attributed to external factors (Richardson, 2011). This perspective underscores the principle that system structure predicates behavior, a notion central to system dynamics and fundamental to understanding complex systems (Sterman, 1994, 2000).

System dynamics (SD) modeling is distinguished by its unique characteristics. These characteristics capture SD's philosophical, theoretical, and practical core (Bala et al., 2017; Richardson, 2011). *Firstly*, SD models are anchored in causal feedback structures, prioritizing causal over correlational relationships within defined problem spaces. *Secondly*, the concept of accumulations and associated delays are fundamental, setting SD apart through stocks and flows that introduce realistic path dependence (Forrester, 1987). *Thirdly*, SD models are equation-based, defining each variable's dynamics through mathematical relationships, thereby facilitating reproducibility and policy analysis. This approach connects stocks, flows, auxiliaries, and constants in a coherent structure that simulates real-world mechanisms. *Fourthly*, SD adopts a continuous time perspective, aligning with real-world processes and enabling detailed feedback loop analysis despite simulations being carried out in discrete time steps (Kampmann & Oliva, 2008).

System Dynamics (SD) has been characterized in various ways within the academic literature. It has been described as a theory (Flood et al., 2003; Jackson, 2006), a method (Coyle et al., 1999; Lane, 2001; Meadows, 1989; Sterman, 2000; Wolstenholme, 2003), a methodology (Roberts, 1978), a distinct field of study (Coyle et al., 1999; Richardson, 2011), a tool (Luna-Reyes & Andersen, 2003), and a paradigm (Olaya, 2009), suggesting its role as a foundational model-building perspective and approach in various disciplines.

Today's system dynamics is essentially interdisciplinary, integrating insights from cognitive psychology, economics, and management to investigate and address complex real-world issues (Sterman, 2000). The discipline has evolved into a theoretical framework and policy design tool, examining how structures within systems manifest observable behaviors and guiding problem-solving across various domains (Forrester, 1987). As Richardson (2011) concisely puts it, system dynamics employs simulation models to identify and analyze the internal drivers of system behavior, facilitating informed policy and decision-making processes.

5 Research Philosophy

As previously highlighted, the significance of research philosophy cannot be overstated due to its profound impact on the perception and construction of reality through different modeling approaches. Moreover, this philosophy plays a critical role in discerning the relative merits of diverse models—a central theme of this dissertation. Consequently, this chapter delves into the evolution of scientific epistemology, illustrating how this progression informs and propels our modeling endeavors.

The philosophical foundations of a discipline are rooted in its core universal assumptions, although identifying these can be challenging given the diversity of perspectives among practitioners and researchers. Despite these differences, it is possible to pinpoint central tenets that link directly to fundamental philosophical queries concerning the nature of reality (ontology) and the basis of knowledge (epistemology). This exploration is essential, as it helps frame the disciplinary inquiry, guiding both theoretical development and practical application. Such a discourse is vital for aligning the discipline's methodology with its epistemological and ontological concerns, thereby shaping its approach to understanding and interpreting the world.

In addition, the validity of research is significantly influenced by its foundational philosophical stance (Cartwright & Montuschi, 2014). Thus, a concise review of the historical evolution of scientific inquiry and its philosophical underpinnings is essential. Consequently, this exploration provides a structured framework to evaluate the validity of the research from a philosophical perspective, ensuring that the chosen approach aligns with established epistemological and ontological principles. This foundational step is critical in delineating the scope, methods, and interpretation of the research findings within the context of its guiding philosophical assumptions.

5.1 From Reductionism to Relativism

Rationalism. The foundations of epistemology can be traced back to René Descartes, who advocated for a methodological shift towards deductive reasoning in philosophy, echoing the precision of mathematical methods. He posited that the only undeniable truths are those revealed through such reasoning, establishing himself as a faithful rationalist. In his work, "Meditations on First Philosophy" (Descartes, 2016), Descartes implemented his "method of doubt" to strip away assumptions, ultimately asserting the certainty of self-awareness through the famous cogito argument: "Cogito, ergo sum" ("I think, therefore I am"). However, while Descartes affirmed the mind's certainty, he maintained a skeptical stance towards the external world, suggesting that while it likely exists, our understanding of it remains perpetually filled with doubt.

Empiricism. Contrasting sharply with Descartes' rationalist approach, John Locke, a proponent of empiricism, set forth a different path in "An Essay Concerning Human Understanding" (Locke, 1847). Locke disputed the existence of innate ideas posited by Descartes, instead portraying the mind at birth as a blank slate, or "tabula rasa," shaped by sensory experiences. Locke's skepticism diverged from Descartes in its foundation; he questioned our knowledge of

the external world based on the inherent limitations of our sensory experiences. For Locke, knowledge begins and ends with experience, meaning that our certainty about the world is continuously shaped and reshaped by our direct interactions with it.

Rational Empiricism. The philosophical landscape of epistemology was further refined by Immanuel Kant, who navigated between the poles of rationalism and empiricism. Kant's work in 1781 was a critical moment in epistemological thought, merging the active mind's conceptual framework from Descartes with Locke's emphasis on experiential input. Kant introduced a critical twist: while sensory experiences trigger ideas, genuine knowledge arises not from passive reception but from the mind's active engagement and structuring of these ideas. Kant introduced categories and forms of intuition that are inherent to the mind, enabling the synthesis of experiences into coherent knowledge. This reorientation suggested that while experiences are crucial, the structure and inherent rules of the mind play a decisive role in shaping our understanding of the world. Kant argued that the mind possesses a priori concepts and categories that structure our understanding, thereby enabling synthetic a priori knowledge - statements that are universally true yet informative about the world (Kant, 1908).

Logical Empiricism. The 20th century witnessed a departure from Kantian epistemology, particularly in the rejection of synthetic a priori knowledge. Logical empiricism, also known as logical positivism, emerged from the Vienna Circle, a group of prominent philosophers who were summoned during the 1920s and 1930s at the University of Vienna. This philosophical movement focused on several key issues: the potential to reduce all synthetic statements to direct observational statements, establishing a strict criterion for meaningfulness, and creating an ideal metalanguage for the philosophical analysis of scientific language systems. In its broadest sense, logical empiricism seeks to delineate meaningful statements into those that are analytically true and those that are empirically verifiable, dismissing statements that fall outside these criteria as metaphysical or meaningless.

The initial iteration of logical empiricism encountered a critical shortcoming regarding the verification principle, especially in the logical substantiation of scientific conjectures. This philosophical stance posited that if a specific outcome, denoted as C, followed from the truth of a theory T, then observing C could confirm T's validity. However, this approach was fundamentally flawed because the occurrence of C could result from an alternative process not covered by theory T, leading to the realization that scientific theories cannot be conclusively verified through direct observation alone.

Karl Popper (1972) addressed this induction dilemma by introducing the falsification principle. He argued that scientific theories should be constructed to be refutable, a shift from seeking irrefutable evidence to embracing the possibility of disproof. Popper maintained that a theory T is considered more credible not when C is observed, but rather when 'not-C' leads to the dismissal of T. This perspective marked a departure from strict verificationism, advocating for a theory's evolution through the accumulation of supporting observations and its immediate rejection upon a single disproof.

Popper's falsificationism softened the firm stance of logical empiricism but did not resolve all its problems. It maintained the division of theories into analytic and synthetic parts and relied

on corresponding empirical observations for each synthetic element. This approach presupposed a clear-cut distinction between verifying instances and their negation. Yet, real-world application revealed the complexity of this ideal, as theories often rest on underlying assumptions, blurring the straightforwardness of empirical validation. Furthermore, the early phase of logical empiricism overly prioritized predictive capability as the sole measure for justifying theories, sidelining the explanatory power and reducing the process to aligning predictions with empirical data. This reduced the role of explanation in scientific inquiry to merely a secondary activity, separate from theory validation, a stance that was increasingly viewed as restrictive and inadequate for capturing the full spectrum of scientific investigation and theory assessment.

Stephen Toulmin (1977) critiques the overemphasis on predictive power as the sole criterion for scientific validation. He suggests that if prediction were the only standard, then even individuals predicting horse race outcomes could be misclassified as scientists, whereas fields such as evolutionary biology, which depend greatly on comprehensive explanatory frameworks, might be unfairly marginalized as unscientific. This observation spurred a recognition among some empiricists of the crucial role of explanation as a form of substantial knowledge. Toulmin points out that this shift necessitates a departure from a purely formalist approach towards a more detailed understanding that encompasses theoretical reinterpretation. This shift suggests that the true value of a scientific theory lies not solely in its predictability but in its capacity to offer insightful and coherent explanations of the observed phenomena, urging a more balanced approach to evaluating scientific theories beyond mere predictive capability (Toulmin, 1977).

Paradigm. Thomas Kuhn, in his work "The Structure of Scientific Revolutions" (1962), revolutionized the philosophy of science by challenging the conventional linear narrative of scientific advancement. He introduced the concept of "paradigm," defining it as the collective mindset that governs the norms, methodologies, and theoretical assumptions within a scientific community. Kuhn argues that science progresses through periods of "normal science," where research operates under an uncontested paradigm, effectively solving puzzles within its scope. However, as anomalies and unsolvable problems accumulate, the foundational beliefs of the existing paradigm come under scrutiny, paving the way for a scientific revolution. This revolution ushers in a new paradigm, fundamentally altering the scientific landscape, including its methodologies, problem definitions, and standards of rationality. Kuhn's perspective suggests that scientific progress is less about an objective march towards truth and more about shifts in community consensus towards more functional and applicable theories. He posits that objectivity is constructed through community agreement rather than innate truth, aligning with relativist viewpoints (Rorty, 1979). Furthermore, Kuhn dismantles the notion of theory-free observation, contending that all scientific observations are inherently colored by the paradigm. This perspective critically undermines logical empiricism and ideas of verification (and by extension, falsification), which assume the possibility of observations untainted by pre-existing theoretical biases. Kuhn's work thus stands as a significant antipositivist critique, reshaping our understanding of the nature and progression of scientific knowledge.

Relativism. The evolution of scientific philosophy towards relativism in the latter half of the 20th century signified a departure from the established logical empiricism that had dominated scientific inquiry. This shift was characterized by a growing critique of the assumption that science could be distilled into a series of formal, context-free analyses. Figures such as Karl Popper played key roles in this transition, with Popper advocating for an appreciation of science's "internal history" while distinguishing between the internal aspects of scientific exploration and the external, contextual influences. Popper's notion of "scientific rationality" sought to balance objectivity with an acknowledgment of the historical backdrop of scientific advancements. Following in Popper's footsteps, Imre Lakatos further diluted the positivist approach, emphasizing the intertwined roles of history and psychology in the fabric of scientific progression. Lakatos critiqued the oversimplified views of falsificationism and stressed the intricacies of theory validation beyond mere empirical evidence (Lakatos, 1970).

By the 1970s, the philosophical and scientific realms began to pivot away from the inflexible frameworks of logical empiricism, opening up to the complexities and practical challenges of contemporary science. This period marked a recognition of the insufficiency of purely formal methodologies to grapple with the complex and multifaceted nature of scientific questions. The incorporation of history, psychology, and sociology into the philosophy of science shifted the discourse towards a more interdisciplinary and refined understanding, moving away from the pursuit of absolute truths towards a focus on practicality and functional utility. As Stephen Toulmin (1977) noted, the fields of history, psychology, and sociology began to play increasingly significant roles in the philosophy of science. Terms such as historicism, relativism, or psychologism, which were once used pejoratively to dismiss philosophical works that incorporated history, sociology, or psychology, were no longer seen in such a negative light. This shift led to a greater openness to interdisciplinary approaches in philosophical discourse. Toulmin observed that the pursuit of absolute, timeless truths had become less fashionable. Practical utility started to take precedence over formal rigor and the traditional notions of "truth" and "excellence." This evolution in epistemology and the philosophy of science signifies that the reductionist and foundationalist philosophy, which had remained largely unchallenged since the 17th century, encountered serious opposition in the form of holistic and relativist philosophy.

5.2 Philosophy of Social Sciences

In social science research, the philosophical orientation adopted by the researcher significantly influences their worldview and methodology, affecting how knowledge is constructed and interpreted. Saunders (2015) outlines principal research philosophies in social sciences, each offering distinct perspectives on the nature of reality and the acquisition of knowledge:

Pragmatism. Pragmatism considers practical effects as vital components of meaning and truth. It suggests that ideas are not static entities but tools for action; their validity is tested through their practical application. It advocates for a practical, problem-solving approach and is not committed to a single system of philosophy or reality (Laudan, 1986). This perspective values diverse research methods and theoretical concepts only insofar as they support practical actions

and outcomes. Pragmatism is versatile, allowing for the integration of different research approaches based on their utility in addressing specific research questions (Bacon, 2012).

Positivism. operates under the assumption that reality is objective and can be described by measurable properties independent of the observer (Comte & Bridges, 2015). Rooted in the natural sciences, positivism emphasizes empirical, observable evidence and the identification of cause-and-effect relationships. It upholds the idea that through systematic methods and statistical analysis, researchers can uncover objective truths about the world (Gauch Jr, 2012; Smith et al., 1996).

Constructivism. Constructivism argues that humans generate knowledge and meaning from an interaction between their experiences and ideas. Unlike empiricism, which states that knowledge is gained largely from sensory experience, constructivism asserts that people construct their own understanding of the world they live in through reflection on experiences (Fosnot, 2013).

Realism. Realism is the belief that reality exists independently of observers. It asserts that objects have an existence outside the mind and that scientific inquiry can reveal truths about them. Realists argue that phenomena should be described as they are, not influenced by emotions, social factors, or personal beliefs. Realism is often contrasted with idealism, which holds that reality is mentally constructed (Niiniluoto, 2017).

Critical Realism. Critical realism combines the concepts of realism and constructivism. It posits that there is a reality independent of human thoughts and beliefs, but our understanding of that reality is always mediated by social conditions and power relations (Bhaskar, 2013). It acknowledges the complexity of the real world, which is often influenced by social, cultural, and historical factors, and seeks to provide a rich, explanatory understanding of how and why social phenomena occur (Niiniluoto, 1999).

Interpretivism. Interpretivism aligns with a subjectivist view, emphasizing the importance of understanding the subjective meanings and lived experiences of individuals or groups (Kroeze, 2012). This philosophy values the context and background of social interactions and phenomena, advocating for a more in-depth, interpretative approach to research that includes the perspectives and interpretations of both informants and researchers (Alharahsheh & Pius, 2020).

Postmodernism. Postmodernism challenges conventional ways of thinking, advocating for a plurality of perspectives and recognizing the fluid, fragmented nature of reality. It emphasizes the importance of understanding individual and collective experiences and deconstructs established narratives and constructs to explore diverse viewpoints and interpretations (Kroeze, 2012).

The choice of research philosophy largely depends on the nature of the research problem and the researcher's objectives. Each philosophy brings different assumptions about the world and dictates how research should be conducted, influencing the choice of methods, the

interpretation of data, and the ultimate conclusions drawn from the research (Crossan, 2003; Holden & Lynch, 2004; Saunders et al., 2015).

Given that trust in automation is a latent construct and its quantification usually requires indirect measurements through questionnaires and/or other behavioral indicators, this dissertation conforms to the assumptions of critical realism. Critical realism acknowledges the complexity of capturing objective reality and endorses the need for estimation in research, a philosophy widely embraced in Human Factors and Ergonomics (HFE). Critical realism position is reflected in how simulation modeling acknowledges that while models represent aspects of real-world systems, they cannot capture all the complexities of those systems. Simulation accepts the existence of a real world that can be partially understood and represented, but also recognizes the limitations and simplifications inherent in models. Critical realism in simulation modeling advocates for a balance between the real structures and mechanisms of systems and the limitations of our knowledge and modeling capabilities.
This dissertation also resonates with principles of Relativism (Interpretivism) in theory building and model development, where the context and background of social interactions and phenomena, encourage an interpretative approach to research. The philosophical underpinning in system dynamics is more aligned with the relativist/holistic philosophy of science. Nevertheless, empirical principles by relying on observational and experimental data to construct and validate models (Empirical validity) are also necessary. Simulations are often used to replicate real-world phenomena in a controlled virtual environment, where empirical data can warrant the simulation processes and outcomes. This approach allows researchers to test hypotheses and observe potential real-world behaviors of systems under different conditions, providing an empirical basis for understanding complex phenomena. Additionally, this dissertation aligns closely with pragmatism, particularly in its focus on the practical application of ideas and the testing of hypotheses for real-world purposes. Pragmatism emphasizes the value of ideas and solutions that are useful in practice (Laudan, 1978). Similarly, simulation models of trust and level of automation can be used to explore 'what-if' scenarios, solve problems, and inform design decision-making processes.

5.3 Research Validity

The understanding of model validity is influenced by the philosophical stance (whether implicit or explicit) on how knowledge is acquired and confirmed. This interplay between different philosophical viewpoints and model validity has been explored by several scholars (Barlas & Carpenter, 1990; Carson & Flood, 1990; Mitroff, 1969; Naylor & Finger, 1967; Senge & Forrester, 1980).

The conventional reductionist/logical positivist approach, which integrates concepts from empiricism, rationalism, verificationism, and strict falsificationism, defines a valid model as an unbiased depiction of an actual system. This perspective insists that models be classified unequivocally as either "accurate" or "inaccurate" and contends that their veracity should become apparent upon comparison with empirical evidence. In this philosophical stance, the primary criterion for validity is accuracy rather than utility (Barlas & Carpenter, 1990). On the other hand, more modern philosophical approaches such as relativism/holism, and pragmatism,

regard a valid model as merely one among several possible representations of reality. From this viewpoint, no single model is inherently superior, although certain models may be more beneficial in particular situations. According to these philosophies, models can never be entirely objective as they inevitably embody the perspectives of their creators. Thus, instead of being judged strictly as true or false, models are appraised based on their practical efficacy along a spectrum of utility (Barlas & Carpenter, 1990).

It is equally important to recognize that the relativist/holistic philosophy, while emphasizing a broader view of model validation, does not dismiss the significance of formal and quantitative tests in this process. Contrary to outright rejection, this philosophy regards validity as something that is gradually established through a process. Within this framework, the act of gathering, organizing, interpreting, and effectively communicating information related to model validity becomes a critical component of the validation process (Barlas, 1996). Formal and quantitative tests are seen as providing essential contributions to the overall validation procedure. This approach underscores the importance of integrating these tests into a larger, more inclusive validation dialogue. Such an approach allows for a more comprehensive and detailed understanding of the model's validity, acknowledging the complexity and multifaceted nature of models and the systems they represent.
Furthermore, in scientific modeling, similar to theories, the definitive accuracy or verification of a model is impossible; a model's comprehensive correctness remains unprovable. The replication of real-world behavior by a model in specific instances does not guarantee its universal accuracy or applicability under varied or future conditions. Instead, a model's validity or its falsity can only be discerned when discrepancies between actual occurrences and simulated outcomes arise, particularly when tested under a diverse array of critical circumstances aimed at falsification. Consequently, discourse in model development shifts from asserting absolute correctness to discussing a model's validity relative to its intended purpose. This relative validity is not a fixed attribute but a provisional status, upheld until contradicted by new evidence.

Bossel (2007) suggests that demonstrating a simulation model's adequacy for representing the real-world system, involves validation across four distinct dimensions: behavioral, structural, empirical, and application validity. *Structural validity* demands that the model's underlying structure—including its state variables and feedback mechanisms—reflects, within the confines of its purpose, the critical structural dynamics of the original system. Descriptive (i.e., correlational) models, lacking a foundational structure, inherently lack structural validity. *Behavioral validity* necessitates that the model replicates the original system's behavior under identical initial conditions and external factors, achieving qualitative concordance in dynamics such as rate changes, oscillation frequencies, peaks, and troughs, phase shifts, equilibria, stability, and responses under extreme situations. *Empirical validity* requires that, within the model's defined purpose, its outputs—whether numerical or logical—align with or are plausible against the real-world system's empirical data under comparable conditions. A model may exhibit behavioral validity without being empirically sound; however, congruence can be achieved by tuning the model with suitable parameters, especially when the system's structure is clear but critical parameters are elusive and necessitate empirical calibration. *Application*

validity entails that the model and its simulation functionalities match the objectives and expectations of its users and the intended application, ensuring that the model serves its designated purpose effectively.

The validation of simulation causal-descriptive models, such as system dynamics, is deeply intertwined with philosophical issues in the philosophy of science. In system dynamics modeling, a model is considered refuted if it can be demonstrated that a relationship within the model contradicts a well-established real-world relationship, even if the model's output behavior aligns with observed system behavior (Pruyt, 2006). For these models, the focus of validation is on the internal structure rather than merely on output behavior – essentially, ensuring the "right behavior for the right reason". A valid system dynamics model, therefore, encapsulates a theory about the functioning of a system in some aspect.
In this context, model validity is not viewed as an absolute concept, and the validation process cannot be entirely objective or formal. Since validity is defined as adequacy with respect to a purpose (Barlas, 2018), model validation necessarily incorporates informal, subjective, and qualitative components. It is seen as a gradual process of building confidence rather than a binary decision of either "accepting" or "rejecting" a model (Lane, 2015).

6 Results and Summary of Appended Articles

This chapter presents the findings of this dissertation, offering analysis and discussion that relate to the background, theoretical frameworks, research methods, and philosophy outlined in previous chapters. It examines how the results relate to established theories and integrates these findings to enhance the comprehension of modeling effort within the context of Human-Automation Interaction (HAI). Key arguments and methodologies used in the associated articles are reviewed and condensed for coherence. As a result, and due to the parallel nature of the content presented here and, in the articles, some duplication of illustrations, tables, and paraphrasing may be observed. Comprehensive details of the results and discussions are fully available in the appended articles in this dissertation.

6.1 Summary and Results of Article 1

This article establishes foundational criteria for assessing models within the Human Factors and Ergonomics (HFE) field, addressing the concerns raised by Dekker and Hollnagel (2004) regarding the scientific validity of theoretical constructs such as situation awareness and trust in automation. It has been argued that these constructs lack theoretical clarity, are unfalsifiable, overly generalized, and rely on descriptive labels rather than explaining causal psychological mechanisms affecting performance (Cass, 2011; Douglas et al., 2007; Flach, 1995; Jodlowski, 2008). To address these concerns, Article 1 focused on HFE as a scientific discipline and developed a set of criteria to assess the scientific credibility of models of trust in automation. Subsequently, this study performed a literature review of the existing models and evaluated the state of the TiA modeling and theoretical progress.

6.1.1 Criteria Development

The criteria development was performed by reviewing the leading scholars and philosophers of science (e.g., Blalock, 1969; Dubin, 1970; Kuhn, 1977; Meleis, 2012; Popper, 1969; Van de Ven, 2007), in combination with Kivunja's (2018) systematic literature review on the fundamental constituents of a scientific theory. Seven fundamental criteria were identified which are briefly described here and summarized with their indicators in Table 2.

Testability or falsifiability, as proposed by Popper, is a fundamental aspect of scientific inquiry, often viewed as the most rigorous standard for evaluating scientific theories (Cramer, 2013). A model's scientific merit hinges on its testability; without the capacity for empirical evaluation, the practical utility of a model remains indeterminate. Testability is generally an empirically driven criterion.

Predictive power. To fulfill the criterion of testability, a model or theory must be capable of generating predictions. Popper (1969) posits that the value of a theory increases with the specificity of its predictions, as these entail a greater risk of refutation, thus enhancing the theory's falsifiability. For instance, the vague assertion that 'A is correlated with B' leaves almost all possibilities open, except for a zero correlation, offering minimal grounds for falsification. In contrast, a more precise claim, such as 'A is positively correlated with B,'

excludes half of the potential outcomes, making the theory more susceptible to empirical disproof. Hence, a model's merit is elevated by its ability to make more explicit and exact predictions.

Explanatory power. The issue with incomplete theories is their ability to predict without providing sufficient explanations for the phenomena observed (Deutsch, 2011). Historically, as noted by Kaplan (1964), ancient astronomers could predict celestial events but lacked the underlying explanatory frameworks. According to Bacharach (1989), a model's utility is dependent upon its capacity for both prediction and explanation, as these are interconnected aspects of robust theorizing. Explanatory models, especially those asserting causal relationships, inherently entail predictions, particularly regarding the outcomes of specific causal actions. Even absent explicit predictions, the narrative of causality typically suggests an expected sequence leading to particular results, underlining the intrinsic link between explanation and prediction in comprehensive theories (Hofman et al., 2017).

Empirical adequacy in scientific theory refers to the accuracy of its assertions regarding observable phenomena (Bhakthavatsalam & Cartwright, 2017; Van Fraassen, 1980). This principle demands that the propositions of the theory align with the observed empirical evidence (Fawcett, 2005). If empirical data supports the model's predictions, then its claims can be tentatively regarded as empirically supported. Conversely, if empirical evidence contradicts the theory's assumptions, the logical inference is that the model's assertions are flawed. It is crucial to distinguish empirical adequacy from empirical testability: the former assesses the truthfulness of a model's predictions based on real-world evidence, while the latter evaluates the model's capacity for being subjected to empirical verification or falsification.

Pragmatic adequacy measures the degree to which a model provides effective solutions to practical problems, reflecting the notion that theories/models are developed for addressing both human and technical challenges to enhance real-world applications (Kerlinger, 1979). In applied fields such as Human Factors and Ergonomics (HFE), this aspect is crucial as the focus lies on translating theoretical insights into practical outcomes. HFE aims to enhance work efficiency, safety, and human well-being, including reducing fatigue and stress while improving overall quality of life (Sanders & McCormick, 1998). Consequently, achieving these objectives necessitates a seamless transfer and integration of knowledge between HFE research and real-world practice (Caple, 2008; Meister, 2018), emphasizing the critical need for theoretical models to be practically relevant and applicable in enhancing human-system interactions.

Human as active agent criterion emphasizes that models should portray humans as proactive entities capable of introspection, decision-making, and the adoption of new concepts and beliefs (Harré, 1984). Accordingly, an effective HFE model should acknowledge humans as active participants and focus on elucidating the underlying mechanisms behind human decisions, behaviors, and perceptions of future occurrences (Kennedy, 2012).

Dynamic properties criterion was formulated in response to the recognition that many HFE issues do not stem from a singular, static cause. Instead, these issues are often emergent properties of intricate interactions within complex socio-technical systems (Guastello, 2017).

Such systems are characterized by their fluid nature, with relationships and connections continuously evolving and adapting within a changing environment (Dekker et al., 2011). Consequently, a critical aspect of evaluation in HFE modeling is the capacity of a model to effectively capture and represent the dynamic behaviors inherent in these phenomena.

Table 2, Model evaluation criteria and their indicators (Poornikoo & Øvergård, 2023)

Criteria	Indicator(s)	Reference
(C1) Testability/ Falsifiability	(1) Can the model be operationalized? Is there a way of measuring the components and constructs in the theory? (2) Does the model/theory propose a research design for testing the model's assumptions? (3) Are the tools and data analysis techniques adequate to measure the model propositions?	Popper (1969), Cramer (2013), Fawcett (1988), Silva (1986)
(C2) Predictive power	Can the model make predictions about: (1) Existence of effect? (2) Direction (or sign) of effect? (3) Direction and interval estimate of effect? (4) Mathematical specification of predicted effect?	Meehl (1967), Dienes (2008), Meehl (1978), Velicer et al. (2008), Freedman (2010), McElreath (2018)
(C3) Explanatory power	Does the model provide: (1) Contrastive force? (2) Explanatory breadth? (3) Explanatory depth?	Cramer (2013), Prochaska, Wright, and Velicer (2008), Garfinkel (1982), Lipton (1990), Ylikoski (2007). Marchionni (2012), Morton (1990), Hitchcock and Woodward (2003)
(C4) Empirical adequacy	(1) Are theoretical assertions made by the model congruent with empirical evidence? (2) Has the entire model been tested in different studies?	Van Fraassen (1980), Bhakthavatsalam and Cartwright (2017), Fawcett (2005), Gould (Gould, 1991), Van de Ven (2007)
(C5) Pragmatic adequacy	Does the model: (1) recognize the domain(s) to which it can be applied to? (2) provides recommendations on how to implement the proposed model in that domain? (3) clarify specific areas in which the model can provide useful and tangible results?	Getty (1995), Karwowski (2005), Caple (2008), Meister (2018), Salas (2008), Sind-Prunier (1996)
(C6) Human as active agent	Does the model take human judgments, motivations, emotions, and socially driven behaviors into consideration?	Witkin and Gottschalk (1988), Gauch (2012), Kennedy (2012)
(C7) Dynamic properties	If the phenomenon is dynamic, does the model acknowledge time as a variable?	Guastello (2017), Dekker, Cilliers, and Hofmeyr (2011), De Keyser et al. (1988)(1988), Hollnagel (2002)

In assessing HFE models (trust in automation in this study), it was crucial to prioritize the evaluation criteria through a hierarchical system due to the varying degrees of importance among different criteria. This study employed the Best Worst Method (BWM) (Rezaei, 2015), a subset of Multi-criteria decision-making (MCDM), which leverages ratios from pairwise comparisons of criteria's relative importance as determined by the evaluator (Liang et al., 2020). The results for each criterion's weight are depicted in Table 3.

Table 3, Pairwise criteria comparison

Criteria Number = 7	Criterion 1	Criterion 2	Criterion 3	Criterion 4	Criterion 5	Criterion 6	Criterion 7
Names of Criteria	Testability	Predictive power	Explanatory power	Empirical adequacy	Pragmatic adequacy	Human as active agent	Dynamic properties

Most important criterion	Testability

Least important criterion	Pragmatic adequacy

Best to Others	Testability	Predictive power	Explanatory power	Empirical adequacy	Pragmatic adequacy	Human as active agent	Dynamic properties
Testability	1	1	5	8	9	9	9

Others to the Worst	Pragmatic adequacy
Testability	9
Predictive power	6
Explanatory power	7
Empirical adequacy	3
Pragmatic adequacy	1
Human as active agent	2
Dynamic properties	3

Weights	Testability	Predictive power	Explanatory power	Empirical adequacy	Pragmatic adequacy	Human as active agent	Dynamic properties
	0,39219046	0,29965114	0,10135259	0,06334537	0,03084644	0,056307	0,056307

6.1.2 Model Evaluation

After an extensive literature review, thirty-six studies were selected for evaluation and categorized into two primary groups. The first group comprises theoretical research aimed at developing conceptual models of trust in automation. These models are often depicted through network diagrams and position trust as a mediating factor influencing the operator's reliance on automation. The second group encompasses computational studies, which seek to formulate mathematical or probabilistic models capable of predicting trust levels. These models integrate various causal factors and explore the relationships between them.

Upon establishing the importance of each evaluation criterion, the assessment proceeds by determining to what extent each model fulfills these criteria. For every criterion, models receive a subjective rating between 1 and 9, which are then normalized as ($X_{norm\,(i,j)}$), and the total scores (OS_i) recalculated according to the formulas:

$$X_{norm\,(i,j)} = \frac{X_{(i,j)}}{Max\ X_j}$$

$$OS_i = \sum (X_{norm\,(i,j)} * W_j)$$

Here, $X_{(i,j)}$ represents the extent to which the ith model meets the jth criterion, X_j is the maximum value in the jth column of matrix X, and W_j denotes the relative significance of the jth criterion.

Additionally, a secondary evaluation with an independent rater was performed on a random selection of 20% of the models (which includes four conceptual and three computational models) to test the reliability of the initial assessments. This step involved calculating the inter-rater reliability to gauge the consistency among different evaluators' ratings, using Krippendorff's alpha (α_k)as the reliability measure. With Krippendorff's alpha value of 0.88, the results indicate a satisfactory level of inter-rater agreement (Krippendorff, 2004, 2011).

The evaluation of models of Trust in Automation (TiA) highlighted distinctive attributes across conceptual and computational models (See Table 4). Among the conceptual models, the Lee and See (2004) framework stands out for its closed-loop dynamic approach and comprehensive consideration of causal factors such as information assimilation and belief formation, alongside individual, organizational, cultural, and environmental contexts. Despite challenges in operationalization and testability, this model significantly contributes to understanding the dynamic nature and foundational dimensions of trust. Desai's (2012) approach, utilizing the Area Under Trust Curve (AUTC), was also a significant advancement in capturing long-term interaction experiences, though it falls short in precise trust prediction and human performance metrics. Kraus's (2020) model which integrates foundational concepts from Lee and See (2004) and Hoff and Bashir (2015), offered a refined view of trust's psychological processes and interactions. On the computational side, models such as Gao and Lee's (2006) EDFT and Hu et al.'s (2019) dynamic human-machine trust models are commendable for their testability, predictive capability, and incorporation of dynamic factors. These models excel in addressing cumulative trust and expectation bias, enhancing predictive power. However, they may lack in accounting for broader causal factors, impacting their explanatory scope and generalizability.

It is unrealistic to expect a model to excel in all proposed evaluation criteria as each model has its strengths and weaknesses across different aspects. This variation in performance among different criteria is why some scholars (e.g., Van Lange, 2013), suggest referring to these benchmarks as 'ideals' rather than strict criteria. The inherent challenge across both model types was balancing detailed operationalization and broad conceptual coverage. Conceptual models provide vital insights but often struggle with precise predictions and operational definitions. In contrast, computational models offer specificity and testability but might miss broader causal relationships. This tension indicates the complexity of TiA modeling, stressing the need for a multifaceted approach that accommodates the dynamic, reciprocal interactions between human agents, automation, and their environment.

Table 4, Normalized Summary Scores of TiA Models (Conceptual and Computational) (Poornikoo & Øvergård, 2023)

Model/Criteria	C1	C2	C3	C4	C5	C6	C7	Overall Score
BWM pairwise weight	0.392	0.300	0.101	0.063	0.031	0.056	0.056	
Muir (1987)	0.63	0.75	0.38	0.75	0.17	0.13	0.13	0.57
Lee and Moray (1992)	0.75	0.75	0.38	0.50	0.50	0.25	0.75	0.66
Muir (1994)	0.75	0.50	0.38	0.50	0.17	0.13	0.25	0.54
Cohen et al. (1997)	0.75	0.50	0.38	0.50	0.67	0.38	0.25	0.57
Madsen and Gregor (2000)	0.38	0.38	0.50	0.25	0.33	0.38	0.13	0.36
Seong and Bisantz (2000)	0.63	0.50	0.38	0.50	0.33	0.13	0.13	0.49
Kelly et al. (2001)	0.63	0.75	0.38	0.50	0.33	0.25	0.13	0.57
Adams et al.(2003)	0.63	0.63	0.75	0.50	0.17	0.38	0.38	0.59
Nickerson and Reilly (2004)	0.50	0.63	0.38	0.25	0.33	0.13	0.25	0.47
Lee and See (2004)	0.75	0.75	1.00	0.75	0.50	0.38	0.50	0.73
Madhavan and Weigmann (2004)	0.50	0.38	0.50	0.75	0.33	0.38	0.25	0.45
Hancock et al. (2011)	0.50	0.50	1.00	1.00	0.33	0.38	0.25	0.56
Desai (2012)	0.75	0.75	0.88	0.50	0.50	0.25	0.38	0.69
Chien et al. (2014)	0.50	0.50	0.88	0.50	0.33	0.38	0.13	0.50
Hoff and Bashir (2015)	0.63	0.63	1.00	0.50	0.33	0.38	0.50	0.62
Bindewald et al. (2018)	0.38	0.50	0.50	0.50	0.17	0.38	0.13	0.41
Kraus et al. (2020)	0.75	0.63	0.75	0.50	0.67	0.38	0.63	0.67
Hou et al. (2021)	0.50	0.38	0.88	0.25	0.17	0.38	0.25	0.45
Gao and Lee (2006)	1.00	1.00	0.88	0.50	0.83	0.63	0.88	0.92
Itoh (2011)	0.63	0.75	0.88	0.50	0.50	0.25	0.25	0.63
Xu & Dudek (2012)	0.88	0.88	0.75	0.50	1.00	0.75	1.00	0.84
Gao et al. (2013)	1.00	1.00	0.88	0.50	0.50	0.38	1.00	0.91
Hoogendoorn et al. (2013)	1.00	1.00	0.63	0.50	0.33	1.00	0.88	0.90
Xu and Dudek (2015)	1.00	1.00	0.50	0.50	1.00	0.38	0.63	0.86
Sadrfaridpour et al. (2016)	1.00	1.00	0.75	0.50	0.67	0.38	0.88	0.89
Akash et al. (2017)	1.00	1.00	0.75	0.75	1.00	0.88	1.00	0.95
Hu et al. (2018)	1.00	1.00	0.88	0.75	1.00	0.88	1.00	0.96
Akash et al. (2018)	1.00	1.00	0.63	0.50	1.00	0.88	1.00	0.92
Chen et al. (2018)	1.00	0.75	0.63	0.50	0.50	0.50	0.88	0.80
Hussein et al. (2019)	0.63	0.63	0.88	0.50	0.33	0.38	0.75	0.63
Sheridan (2019)	0.63	0.63	0.88	0.50	0.33	0.38	0.38	0.61
Nam et al. (2020)	1.00	1.00	0.75	0.50	1.00	0.88	0.75	0.92
Guo and Yang (2020)	0.88	1.00	0.63	0.50	0.83	0.88	1.00	0.87
Chen et al. (2020)	0.88	1.00	0.75	0.50	0.67	0.63	1.00	0.86
Azevedo-Sa et al. (2021)	0.88	1.00	0.63	0.50	1.00	0.50	0.88	0.85

The essence of this study was to refine the model evaluation process in HFE, specifically applied to trust in automation (TiA), illustrating the divergence between conceptual and computational models. This study suggested that while the established criteria draw from the philosophy of science, adaptations may better serve specific HFE contexts. The use of the Best-

Worst Method (BWM) for criteria ranking introduces a subjective layer; hence, incorporating insights from Subject Matter Experts (SMEs) could mitigate bias and enhance the robustness of future evaluations. Additionally, aligning evaluation criteria with specific application contexts could guide selecting the most appropriate models for theoretical exploration or practical design, fostering a more tailored and effective approach to HFE research and application. This analysis also underscored the stagnation in TiA research progression, signaling the HFE community's challenge with trust's complexity and the limitations of conventional modeling approaches. To address this, adopting more advanced modeling techniques such as system dynamics and agent-based modeling was recommended to provide a richer, more comprehensive understanding of trust dynamics within sociotechnical systems.

6.2 Summary and Results of Article 2

Article 2 discusses the challenges and complexity of modeling human behavior in sociotechnical systems, which are subject to varying technical and contextual factors. The complexity of these systems makes providing a reliable explanation of human behavior remarkably difficult, particularly in the context of human-automation interaction (HAI), which stands as a prominent issue in cognitive engineering today.

Article 2 further elaborates on Human Factors and Ergonomics (HFE) as a discipline grounded in systems theory, examining its systemic characteristics. It delves into the epistemological appropriateness of various modeling approaches for addressing the complexities inherent in systems.

6.2.1 HFE As a System Discipline

Human Factors and Ergonomics (HFE) fundamentally constitutes a systems-oriented discipline. It encompasses various systems, including socio-technical and cognitive systems, aiming to improve both performance and well-being through design enhancement and better human-system integration. Wilson (2014) identifies six key systemic characteristics (Figure 10) inherent to HFE: system focus, context, interactions, holism, emergence, and embedding. *System focus* underlines the significance of examining the network of interactions among various elements, whether organic or inorganic, across different levels of functionality and conceptual frameworks. *Context* acknowledges that behaviors and performances are inherently influenced by their environments, challenging the generalizability of laboratory findings to complex real-world scenarios. *Interaction* emphasizes the essential nature of relationships within systems, vital for optimizing human and technological collaborations. *Holism* advocates for a comprehensive approach, considering physical, technical, cognitive, and social facets to craft effective solutions. *Emergence* recognizes that systems may exhibit properties unforeseen by their designers, highlighting the adaptability and creative potential of users. Finally,

embedding emphasizes the participative aspect of HFE, incorporating insights from stakeholders and experts into ergonomic practices.

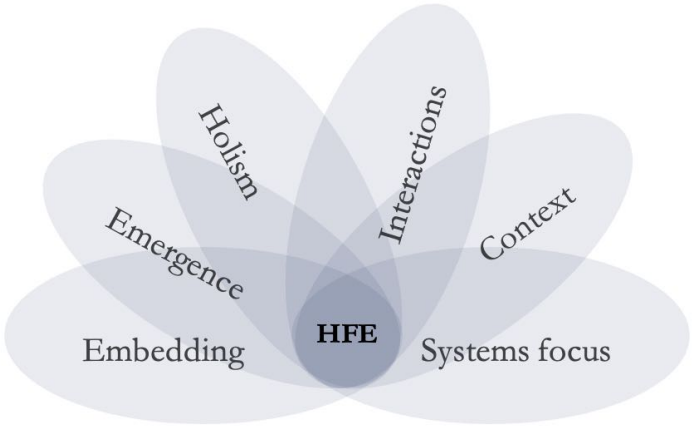


Figure 10, HFE System Characteristics adopted from Wilson (2014)

Given these premises, it becomes apparent that HFE models should represent interconnected networks operating within closed-loop systems, embodying systems thinking. This perspective necessitates an examination of the epistemological underpinnings of various HFE modeling approaches to determine if they accurately reflect systems thinking in their methodological frameworks.

6.2.2 Epistemological Assumptions of Modeling Approaches

In exploring modeling issues related to human interactions within automated systems, two primary epistemological approaches have been identified. Bruner (1986) distinguishes these based on differing thought modes: the *paradigmatic or logico-scientific mode*, which focuses on variance, and the *narrative or process mode*, which emphasizes storytelling and the sequence of events. These approaches offer distinct perspectives on reality construction and validation methods. Aldrich (2001) further categorizes these into outcome-driven approaches, answering 'what' questions, and event-driven approaches, addressing 'how' questions, each providing unique insights into HFE problems and solutions. Article 2 adopts this perspective for the suitability of modeling approaches for studying Human-Automation Interaction (HAI).

The outcome-driven, or variance approach (Van de Ven, 2007) focuses on elucidating the relationships between independent and dependent variables to tackle 'what' questions, such as identifying the antecedents and consequences related to a specific phenomenon. This approach demands evidence of co-variation, temporal precedence, and elimination of spurious relationships between variables. Employing research designs such as experiments and surveys, variance models rely on the general linear model framework, supporting the use of statistical analyses including ANOVA, regression, factor analysis, and structural equation modeling to validate hypotheses and uncover patterns within HFE research.
Contrary to outcome-driven variance models, process models focus on event-driven explanations, addressing the "How" question. They seek to explain the sequence of events by

revealing the mechanisms responsible for causing real-world occurrences and the particular conditions or contingencies that activate these mechanisms. This approach aims to understand the dynamic and complex interactions that lead to specific outcomes, providing a more reflective understanding of the processes at play.

The epistemological assumptions in variance models present a specific method for interpreting reality, segmenting it into analyzable and quantifiable segments. The variance approach is particularly adept at exploring questions that involve comparing different entities or understanding linear causal connections between variables. Nonetheless, this approach encounters limitations when applied to the examination of social entities, especially within the context of sociotechnical systems, where dynamics are complex and multidimensional.
On the other hand, the assumptions in process models feature a different outlook, where causes are seen not just as immediate triggers but as parts of a continuous interaction within the system's history. Particularly, such models emphasize the emergence of causal forces from specific feedback mechanisms, emphasizing the temporal and evolving nature of cause-effect relationships. The fundamental assumptions of the two modeling perspectives are described in Table 5.

Table 5, Outcome- and Event-driven Models

Outcome-driven variance model	Event-driven process model
The universe is composed of stable entities with attributes that vary.	Entities engaged in events are dynamic, subject to change over time rather than being static.
Explanations are fundamentally rooted in the concept of efficient causality.	Explanations draw on a mix of final, formal, and efficient causality, offering a multifaceted perspective on causes and effects.
The universality of any given explanation is contingent upon its applicability across diverse scenarios.	The value of explanations lies in their adaptability and applicability across different contexts and conditions.
The order of occurrence of independent variables relative to dependent variables does not impact the final outcomes.	Chronology is crucial; the order in which events unfold significantly impacts outcomes.
Explanations should prioritize immediate or direct causation.	A comprehensive explanation should consider a spectrum of influences, from the proximate to the more remote.
Attributes retain a singular causal interpretation over time.	The significance of an entity, attribute, or event can evolve, reflecting its changing role and impact over time.

One important consideration is understanding the interplay between the two approaches and their complementary nature. The rationale within a variance model, whether stated directly or indirectly, outlines the narrative of how specific conditions lead an independent (input) variable to influence a dependent (outcome) variable. Thus, delving into the underlying process believed to justify the causal relationship between independent and dependent variables can enhance the

solidity of responses to 'what' (variance theory) questions. Conversely, responses to 'how' questions may lack significance without addressing the corresponding 'what caused it?' or 'what are its consequences?' questions, underscoring the necessity of integrating both variance and process theories for a more comprehensive understanding.

If we accept Human Factors and Ergonomics (HFE) and its sub-disciplines including HAI, as a systems discipline, it is crucial to identify the types of models that effectively facilitate the study of human performance within such sociotechnical systems. Variance models, with their analytical reductionism, fall short of capturing the dynamic interplay and mutual impacts among different system elements when confronted with simultaneous multiple impacts, due to the complexity inherent in modern high-technology systems. This complexity, a hallmark of today's systems, emphasizes the importance of examining systems holistically, focusing on interrelationships rather than isolated components. In such systems, behaviors emerge that are not predictable from the properties of individual parts. Despite the prevalence of variance-based methodological approaches in HFE, which facilitate the study of individual operators, teams, and technical performance through established toolkits and methods, the bidirectional and nonlinear nature of cause-effect relationships within sociotechnical systems calls for a shift towards more integrative process models. These models, though more complex to develop, are better suited for capturing the multifaceted interactions, feedback loops, and temporal dynamics inherent in social entities and human-machine interactions.
Article 2 concludes by emphasizing the complementary nature of variance and process approaches, where HFE should leverage the existing consensus on causal factors within variance models to inform the development of more advanced process models. By embracing a systems approach and focusing on the complexities of modern sociotechnical systems, HFE and HAI research can evolve towards providing more comprehensive causal models that address both 'what' and 'how' questions, thereby deepening our understanding of human performance in sociotechnical environments and advancing the formalization of HFE theories.

6.3 Summary and Results of Article 3

Utilizing a fuzzy rule-based inference system paired with operational criteria for automation, Article 3 offers quantification and definition of LOAs, particularly tailored to operational tasks and functions within Maritime Autonomous Surface Ships (MASS). One of the major outcomes of this article is to establish a standardized language to articulate LOAs in MASS, thereby transforming an abstract construct into a concrete, operationalizable concept.

To develop Levels of Automation (LOAs), this article focused on LOAs at the function and system levels which adopt the operational criteria by Parasuraman et al. (2000) and recommended by Veritas (2019) for autonomous shipping. These guidelines classify automation's support to humans into four stages: information acquisition, information analysis, decision selection, and action implementation, where each stage can take various degrees of automation, as shown in Figure 11.

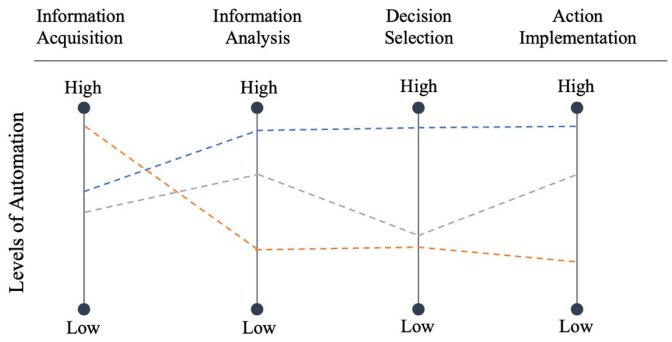


Figure 11, Operational criteria for levels of automation, adapted from Parasuraman et al. (2000)

In the next step, utilizing fuzzy logic, *truth* is defined as a quintuple $(H, T(H), U, G, M)$, where H is a variable's designation; $T(H)$ denotes a finite set of linguistic values for H; U symbolizes a universe of discourse; G represents a set of rules for generating $T(H)$; and M stands for a membership function mapping terms in $T(H)$. In this context, LOAs and their linguistic term sets can be formulated as: $T\ (LOAs)\ =\ Very\ Low, Low, Medium, High, Very\ High$.

The fuzzy logic steps unfold in four primary stages, as illustrated in Figure 12.

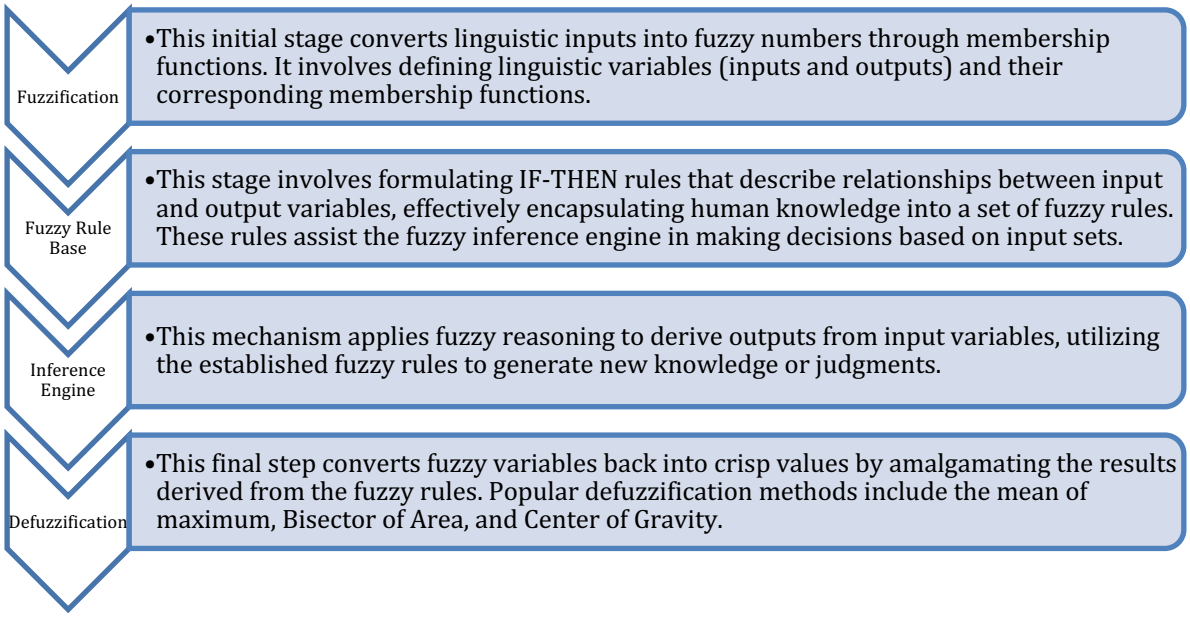
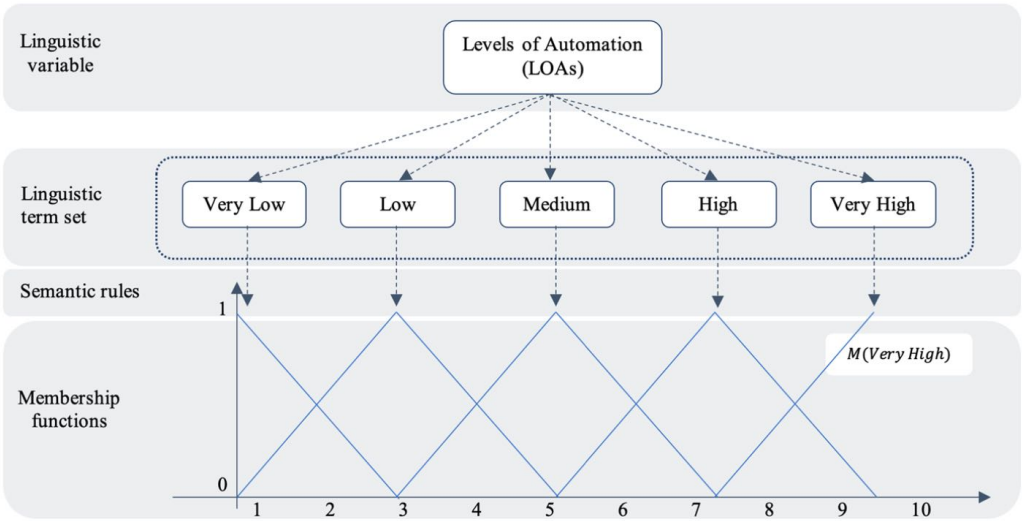


Figure 12, Fuzzy logic steps.

Using the MATLAB fuzzy logic toolbox, the process begins with the establishment of fuzzy inference systems (FIS), selecting Mamdani for the Fuzzy Inference System (FIS) type and the Centroid method for defuzzification.

In the fuzzification step, membership functions for input variables, which in this case are information acquisition, information analysis, decision selection, and action implementation, were defined. According to established best practices, selecting three to seven linguistic term sets for these variables is advisable to maintain clarity and manageability. In this specific article,

three linguistic terms are chosen for input variables to maintain simplicity and efficiency without compromising the system's effectiveness. In contrast, five linguistic terms are selected for the output variable to allow for a more accurate representation of LOAs. A graphical



depiction of this process is illustrated in Figure 13.

Figure 13, Process of defining membership function of LOAs linguistic terms (Poornikoo & Øvergård, 2022)

The Gaussian membership function is utilized for both input and output variables, an approach often chosen due to its effectiveness in handling nonlinear transitions between different automation levels. This function's shape is well-suited for representing the uncertainty and ambiguity inherent in natural language terms, thereby providing a smooth transition between different levels of automation. The parameters for these membership functions, which define their shape and spread, are detailed in Table 6. These parameters are crucial as they influence how input data is interpreted and classified into different linguistic terms. Figure 14 shows the membership functions associated with the input and output functions. For any given input value, one or more membership functions can be engaged in the operation. For example, input variable=3 for information acquisitions may activate membership functions 'Low' and 'Medium'.

Table 6, Inputs, and Output membership functions type and parameters (Poornikoo & Øvergård, 2022)

variables	Type	Term 1, Parameters	Term 2, Parameters	Term 3, Parameters	Term 4, Parameters	Term 5, Parameters
INPUT1	Gaussian	Low [1.699 6.939e-17]	Medium [1.699 5]	High [1.699 10]		
INPUT2	Gaussian	Low [1.699 6.939e-17]	Medium [1.699 5]	High [1.699 10]		
INPUT3	Gaussian	Low [1.699 6.939e-17]	Medium [1.699 5]	High [1.699 10]		
INPUT4	Gaussian	Low [1.699 6.939e-17]	Medium [1.699 5]	High [1.699 10]		
OUTPUT	Gaussian	LOA 1 [1.062 -2.776e-17]	LOA 2 [1.062 2.5]	LOA 3 [1.062 5]	LOA 4 [1.062 7.5]	LOA 5 [1.062 10]

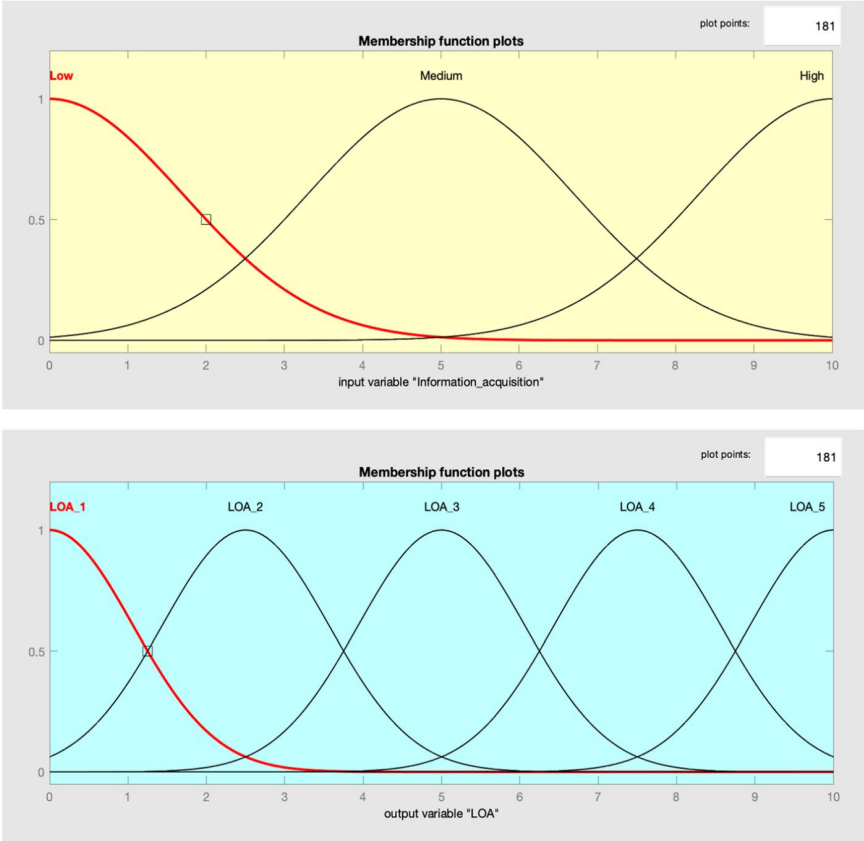


Figure 14, Inputs, and Output Gaussian membership functions (Poornikoo & Øvergård, 2022)

By applying IF-THEN rules, the Fuzzy Inference System (FIS) determines levels of automation (LOA) based on input variables. The FIS integrates multiple input variables—information acquisition, information analysis, decision selection, and action execution—into a cohesive framework that defines the automation level. By crafting 81 fuzzy rules, FIS maps these inputs onto the outcome variable LOA. Finally, the outcome of the model is defuzzified to return crisp values for LOA. The simulation model can effectively take any values as inputs and return the function's LOA, as illustrated in Figure 15. The returned crisp values of LOAs denote automation levels at a functional level. A similar process of fuzzification can be implemented to identify the LOAs of a system (Figure 16).

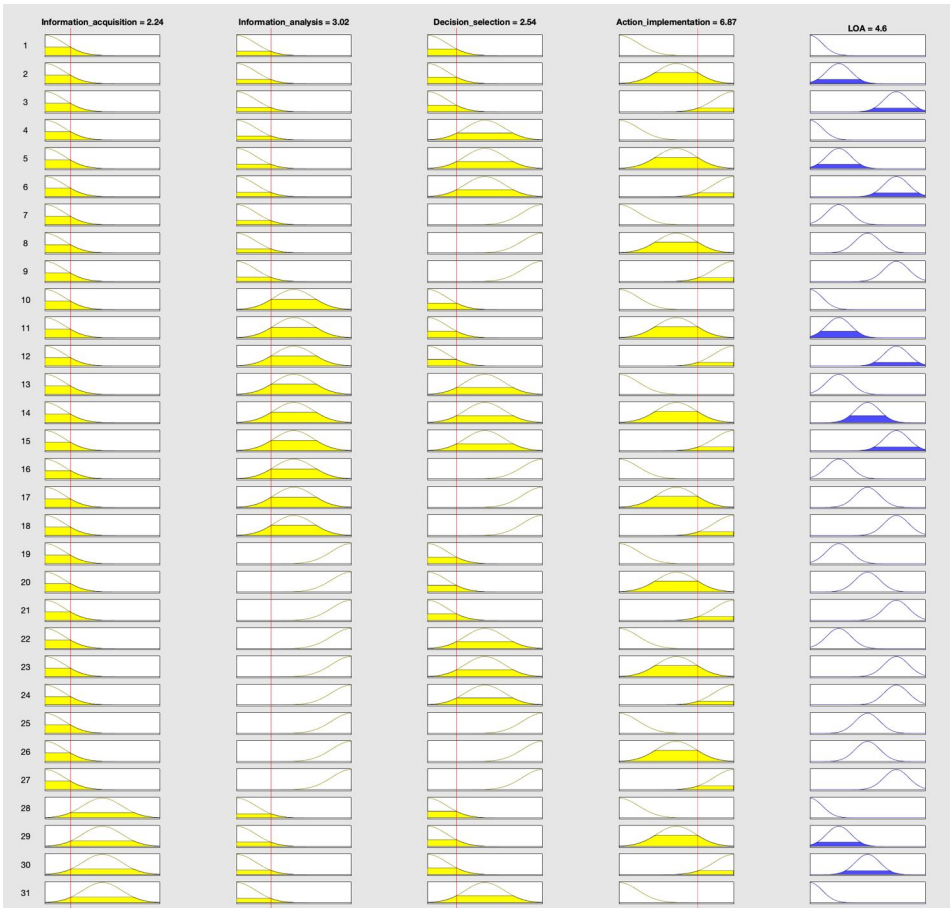


Figure 15, Rule viewer for 4 inputs and output variables (Poornikoo & Øvergård, 2022)

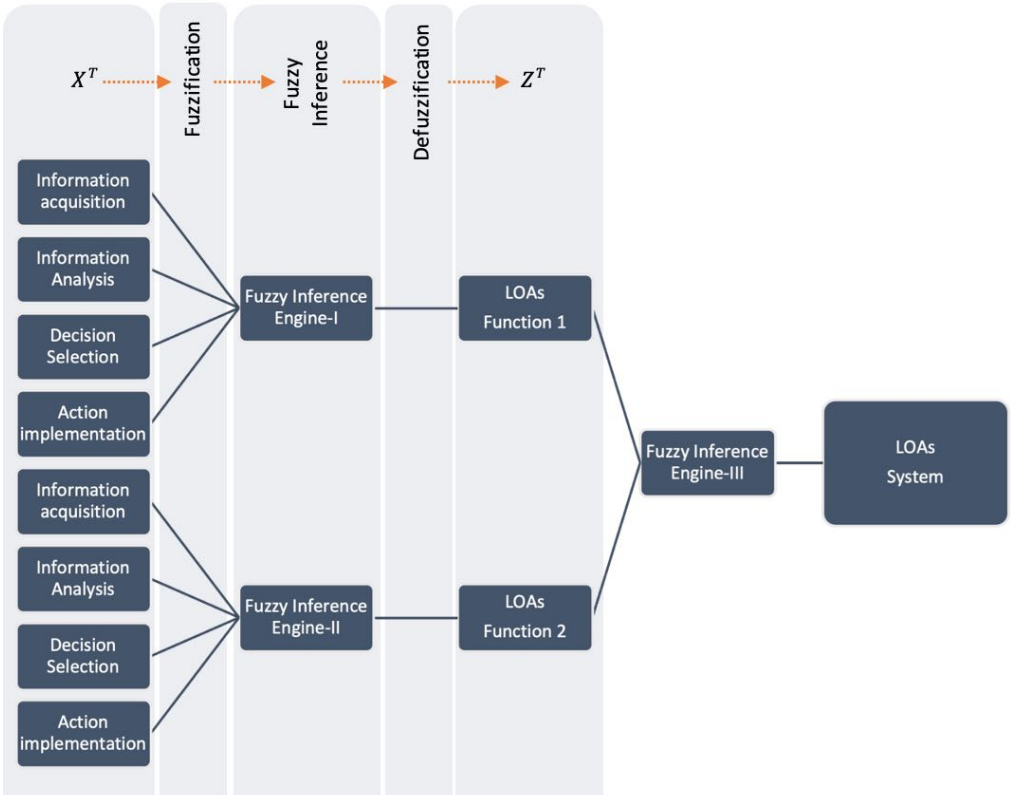


Figure 16, Fuzzy LOA process across tasks, functions, and system (Poornikoo & Øvergård, 2022)

Utilizing a bottom-up hierarchical framework combined with fuzzy logic, this article developed a structured approach for determining the autonomy levels in various tasks, functions, and systems. The simulation outcomes demonstrate the adaptability of fuzzy membership functions to cover a broad spectrum of autonomous activities, thus defining the Levels of Automation (LOA) for autonomous vessels effectively. The primary benefit of this proposed method lies in its practical application to an otherwise abstract concept typically used to outline the potential capabilities of MASS. The proposed fuzzy logic approach not only characterizes an autonomous vessel based on its fundamental functions but also enables a dynamic representation of how automated each function is. For example, the level of automation in navigation can be adjusted based on the required human interaction or lack thereof. This adjustment necessitates a detailed task analysis, particularly for the navigation function and its dependencies across various operational contexts. In this model, tasks such as information gathering and analysis could be assigned high autonomy levels (i.e., membership functions nearing one) due to the ship's reliance on advanced sensory technologies including optical and infrared cameras, LiDAR, and RADAR. However, in scenarios of low visibility or high traffic, human intervention may still be essential for decision-making and action execution, which indicates the need for dynamic LOA adjustments in real-time navigation.

6.4 Summary and Results of Article 4

Article 4 builds upon articles 1 and 2 while addressing some of the key limitations of Trust in Automation (TiA) models including:

1. Trust dynamics: most previous research has treated trust as a static attribute, captured at a single moment, failing to acknowledge its evolving nature due to ongoing interactions with automation.
2. Conceptual model limitations: many conceptual models use broad, general terminologies, which limit their testability and empirical validation. They often lack specificity in predicting how trust changes in response to different factors, making them less actionable for real-world applications.
3. Computational model constraints: computational models, while providing numerical insights into trust variations, are often too tied to specific datasets, limiting their generalizability across different contexts of human-automation interaction.
4. Practical relevance: the practical utility of existing TiA models is limited due to their shortcomings in identifying specific intervention points for adjusting trust levels. This restricts their usefulness in designing and implementing automation systems that foster appropriate levels of trust.

To address these issues, article 4 introduces a system dynamic simulation model aimed at exploring structural aspects of Trust in Automation (TiA). The model accounts for the phases of trust development, decline, and recovery during user interaction with automated systems.
By employing a System Dynamics (SD) methodology, this model demonstrates the complex, nonlinear interplay of trust via dynamic feedback mechanisms. Contrary to other methodologies that depend heavily on empirical data for system behavior, SD emphasizes the development of formal models that encapsulate dynamic phenomena as continuous feedback processes. These

conceptualizations embed propositions about the cause-and-effect dynamics among elements and variables within the system, analyzing the outcomes of their interplay. In this schema, feedback loops are considered analytical units, each designed for a specific function and exhibiting variable importance over time. Variables may participate in several feedback loops, shifting the focus from isolated causal links between variable pairs to a more extensive examination of causative frameworks.
The model accounts for key variables responsible for trusting behavior including system performance (reliability), system capabilities, system malfunctions, perceived reliability, expectations of performance, perceived risk, and operator's individual characteristics.
Trust levels can fluctuate in response to the system's performance (Lee & Moray, 1992; Muir, 1994), with such trust variations often aligning positively with changes in automation utilization. A decline in trust prompts operators to favor manual operations over automated solutions, reducing their willingness to explore the automation's functions. Continuous interaction with the system also cultivates operators' performance expectations for automation, aligning with the notion of "anticipation" in Sheridan's (2019) framework. Moreover, prolonged engagement fosters generalized perceptions of automation efficacy and broader constructive beliefs about system conduct, encapsulated as "faith".

The outcome of the model is (partially) assessed through an experimental study involving participants engaging with simulated Maritime Autonomous Surface Ships (MASS).

6.4.1 Model Structure

Figure 17 illustrates a simplified TiA Causal Loop Diagram (CLD). The dynamics of trust in automation (TiA) are captured through the interplay between three key components: Trust (T), Perceived Performance, and Expectation of Performance. The central component, Trust (T), changes over time based on the interactions between Perceived and Expected Performance.
Assume $T(t)$ be the *Trust* stock at time t and $c(t)$ be the *Change in TiA*. Using this convention, the following formulation can be articulated, where ρ denotes the *Initial Trust*, θ denotes the *Expectation Gap*, ω indicates the *Difference Between the Maximum TiA (Faith)* and *Current Trust*, and λ represents the *Trust Adjustment Time*.

$$T(t) = \rho + \int c(t)dt \quad (1)$$

$$c(t) = \min\left\{\frac{\theta}{\lambda}, \frac{\omega}{\lambda}\right\} \quad (2)$$

Initial Trust (ρ) is a reflection of an individual's baseline level of trust before interacting with the automation. It is influenced by previous experiences and inherent trust propensity (Merritt & Ilgen, 2008). This means that two individuals might start from different levels of trust based on their past interactions with similar systems and their general inclination to trust technology. This variability is referred to as intra-individual variability.
Adjustment Time refers to the rate at which individuals update their trust in response to new information or experiences. This rate can vary significantly among individuals. Some people may quickly reassess their level of trust after a single event (fast adjustment time), while others

may require more time or additional evidence to alter their trust levels (slow adjustment time). These differences are crucial as they influence how each person's trust evolves in response to the perceived performance of the automation.

Despite the individual differences in initial trust levels and adjustment times, the model assumes that the basic feedback mechanisms underlying trust dynamics are common across individuals. This implies that while the starting point (initial trust) and the rate of change (adjustment times) can differ, the way in which trust increases or decreases in response to the gap between expected and perceived performance follows a universal pattern. This approach allows the model to accommodate individual differences while maintaining a consistent structure for understanding how trust evolves over time. This combination of individual variability with a standard feedback structure enables a more personalized understanding of trust dynamics, catering to the complex nature of human-automation interactions. By considering individual characteristics, the model can provide more accurate predictions and insights into how different people will react to the same system, which is critical for designing, implementing, and managing automated systems effectively.

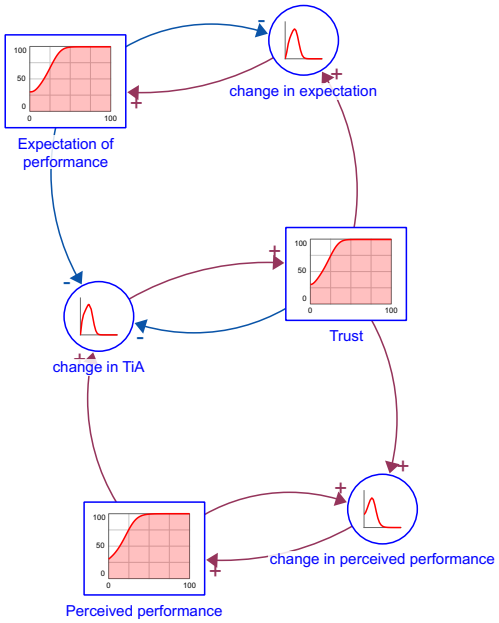


Figure 17, Simplified TiA Causal Loop Diagram (CLD)

The system dynamics model encompasses three main feedback loops (as shown in Figure 18) influencing trust dynamics in autonomous systems: a reinforcing loop (R1) where increased trust leads to more trusting behavior, moderated by perceived risk; a balancing loop (B1) where trust adjusts according to the gap between expected and perceived performance, thus ensuring trust reflects actual system performance; and another balancing loop (B2) that caps trust growth at a maximum threshold (*Faith*) to prevent unlimited trust. Additionally, the model factors in system malfunctions and systems capabilities, reflecting how sudden failures or inherent system limitations can impact trust levels. Figure 19 illustrates the basic Stock and Flow Diagram (SFD) of the model structure. System malfunction in this model represents a sudden disruption in the System performance, formulated with a Pulse function that incorporates the *Magnitude*

and *Time* of the System Malfunction. On the other hand, System Capability reflects the limiting cap to the Perceived Performance. Environmental Challenges, such as adverse weather conditions or limited visibility, can greatly impact the System Capability. A more thorough description of the model structure can be found in the appended articles.

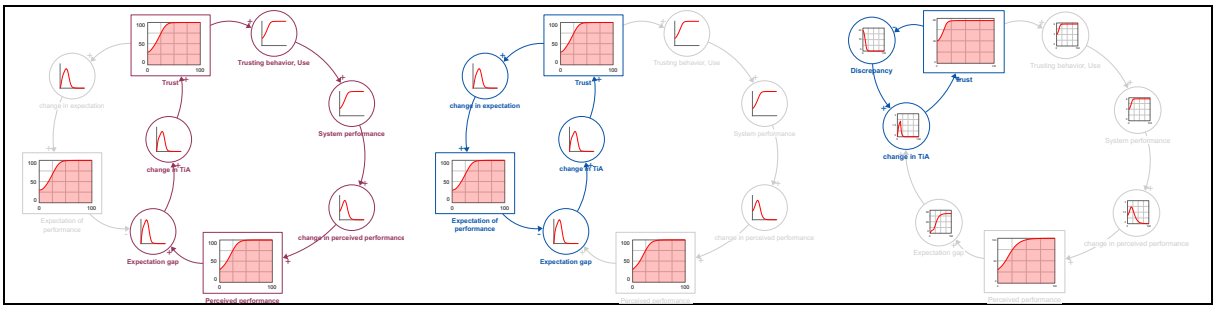


Figure 18, Model's three main feedback loops.

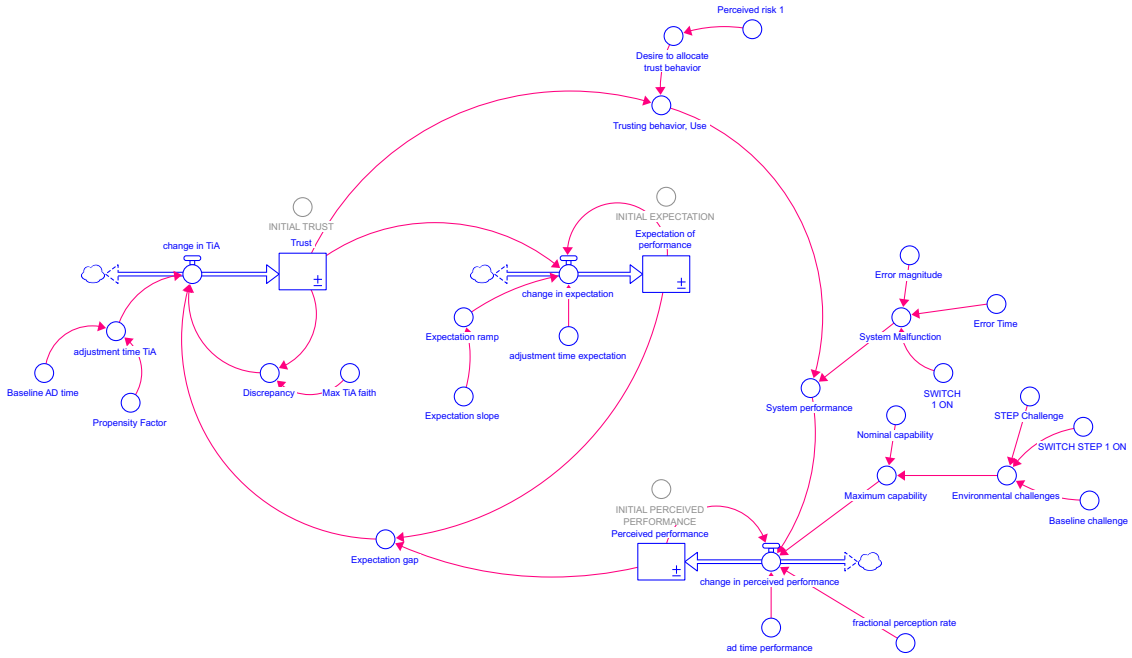


Figure 19, Stock and Flow Diagram (SFD)

6.4.2 Simulation Results

The simulation results highlight various dynamics in trust in automation. In a baseline scenario (i.e., perfect automation), the model stays at equilibrium when the system's capabilities match the operator's initial trust (for example, 30 out of 100) (see Figure 20). Deviations from this state trigger the reinforcing loop (R1), leading to different trust behaviors based on the initial trust levels, showcasing goal-seeking and path-dependency behaviors where past states influence future behaviors as illustrated in Figure 21. This process indicates that operators with higher trust levels tend to engage more with the system to evaluate its performance, resulting in an increased positive perception of performance, assuming the automation is flawless. This kind of path-dependent trust evolution has been observed in various studies (Castelfranchi & Falcone, 2010; Lewis & Weigert, 2012), indicating how past and initial trust conditions significantly influence future trust levels and actions. The presence of path dependency in trust

dynamics is further affected by which feedback loops are predominant at any time (i.e., loop dominance). As circumstances evolve, different loops may gain prominence, directing the pace and direction of trust development. This suggests that the system's behavior is influenced not only by its starting state but also by the interplay and dominance of various feedback mechanisms over time, leading to intricate and occasionally unexpected trajectories of trust progression.

The baseline simulation findings reveal a notable path dependency in the evolution of individual trust trajectories. As previously discussed, the divergence in individual experiences is captured through variations in *Initial Trust* values and the time it takes for individuals to adjust their trust levels. The model highlights the dynamics of trust formation and erosion by showcasing the speed at which trust is gained or lost, a concept described as the gradient of trust dynamics. These intra-individual variations are illustrated in Figure 25, demonstrating that while individuals may start with different Propensity to Trust, their trust development trajectories exhibit similar patterns. Additionally, the model points to a convergence towards a state of path dependency, influenced by different Initial Trust values, as depicted in Figure 22. This convergence is a result of adjustments made in response to the gap between expected and actual performance. When the perceived performance aligns with or surpasses expectations, there is a gradual increase in trust.

The variability in trust-related outcomes can be significantly impacted by the initial discrepancy between Expected Performance and Perceived Performance. When the performance anticipated by an individual far exceeds the performance actually observed, this gap triggers a negative feedback loop, resulting in an initial reduction in trust towards automation. This decrease in trust leads to a more cautious approach towards engaging with the automation system, causing a slower recognition of its effectiveness. This results in a prolonged phase of trust building and adjustment of expectations. As the difference between what is expected and what is perceived decreases, trust starts to build up again, though at a slower rate than initially might have been the case, as illustrated in Figure 23. This demonstrates how early expectations and subsequent performance perceptions are crucial in shaping the trajectory of trust development in automated systems.

To delve deeper into the dynamics of trust in the context of imperfect automation, a disruption in the baseline model was introduced by implementing a Pulse function. This function models an unforeseen error with a magnitude of 50% occurring at the specific moment of t=25, resulting in an immediate and significant drop in both System Performance and Perceived Performance, as depicted in Figure 24. The introduction of this error creates a gap between expected and actual system performance, triggering a delay in the reduction of trust. This reduction sustains trust at a diminished level until system performance begins to recover and realign with the operator's expectations. The impact of a system malfunction on trust does not manifest instantly but rather initiates a gradual erosion of trust over time. Recovery from such malfunctions also does not happen immediately but occurs gradually, as highlighted in Figure 24, where trust ascends to a level lower than that of the perceived performance improvement post-malfunction (Yang et al., 2017).

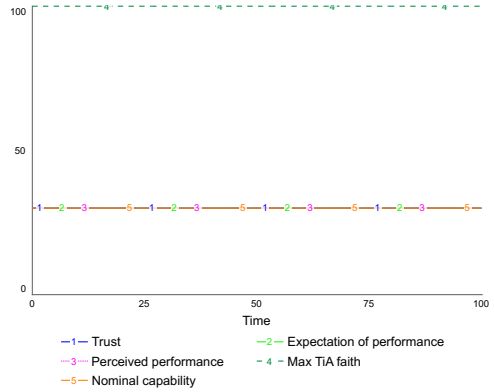


Figure 20, Model at equilibrium

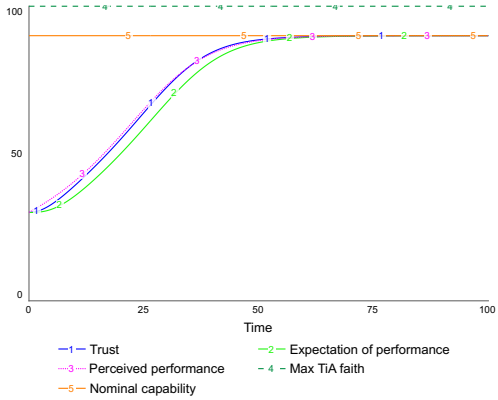


Figure 21, Model's S-shape growth & path dependency

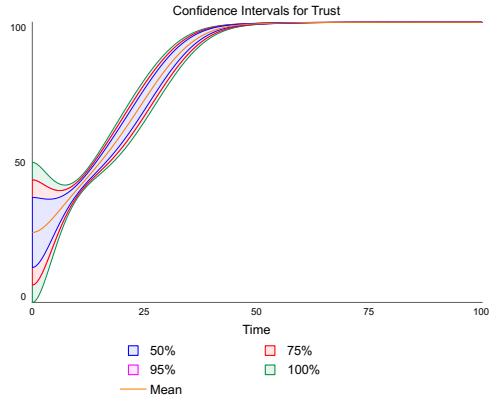


Figure 22, Individual variability in trust evolution (initial trust)

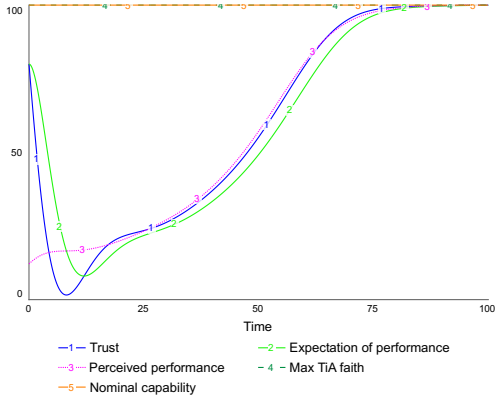


Figure 23, Mismatches between expected and perceived performance

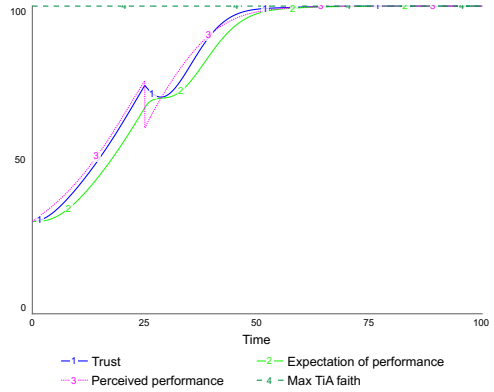


Figure 24, Trust decline as a result of system malfunction.

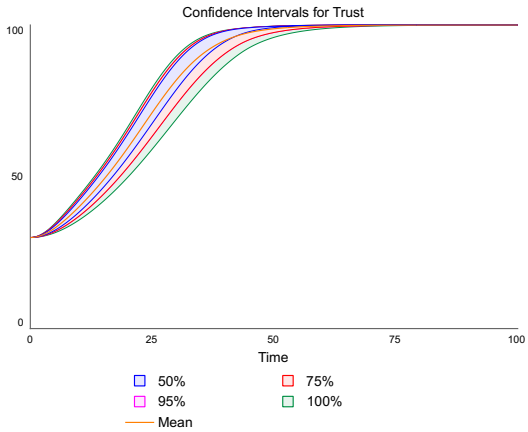


Figure 25, Individual variability in propensity to trust

To assess the impact of system malfunctions occurring at different stages of the interaction, multiple simulation runs were investigated. As delineated in Figure 26, the timing of these errors critically affects the trajectory of trust recovery post-malfunction. The uppermost trajectory in the figure, which shows no decline, illustrates the baseline scenario where the

system operates without encountering any errors. Introducing system malfunctions early in the interaction, approximately at time t=10, leads to a significant initial drop in trust. This early disruption results in a lengthy recovery phase, during which trust levels fail to return to the baseline (i.e., perfect automation) within the observed period. Conversely, malfunctions that occur midway through the interaction, around time t=25, precipitate a noticeable but less severe drop in trust, with a subsequent recovery phase that is shorter compared to early errors. Malfunctions occurring later in the interaction process have the least impact on trust levels, with trust recovering more rapidly and nearly reaching the levels observed in the baseline scenario. This analysis illustrates the complex relationship between the timing of system errors and the resilience of trust in automation. It highlights how earlier errors can severely disrupt the trust-building process, necessitating extended periods for trust to be re-established, while errors occurring later are less detrimental, allowing for a quicker restoration of trust.

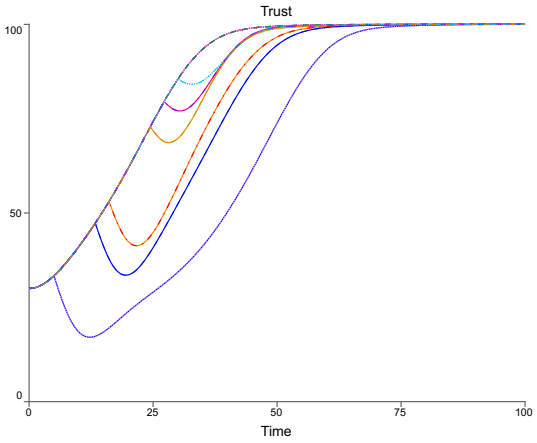


Figure 26, Multiple simulation runs with varying error time

6.4.3 Empirical Study

Article 4 also presents empirical findings from an experiment designed to validate a segment of the proposed System Dynamics model, specifically tailored for application in Maritime Autonomous Surface Ships (MASS). The motivation behind selecting a portion of the model for empirical validation stems from the uncertainty surrounding the time required for an individual to attain maximum trust levels, a parameter challenging to test directly in practice. However, since the model addresses the decline in trust, a phenomenon that can occur more rapidly and is thus amenable to experimental scrutiny and validation. The secondary goal was to explore the behavioral manifestations of trust in automation by analyzing operators' eye movements during monitoring tasks. This includes examining how often participants monitor the system and the characteristics of their fixations to understand the impact of system malfunctions on their engagement.

A. Participants

The experiment involved 30 participants, including nautical students and instructors from a maritime university. These participants, recruited through snowball sampling, ranged in age from 18 to 55 and included a mix of 22 males and 8 females.

B. Apparatus

For this study, the Kongsberg K-Sim desktop bridge training simulator (Kongsberg, 2023) served as the testing ground, simulating maritime conditions for autonomous vessel operation. The setup involved three main components: a RADAR display, a central bridge control interface, and an Electronic Chart Display and Information System (ECDIS), as depicted in Figure 27.

Figure 27, Experiment Setup; Navigation Lab, University of South-Eastern Norway (USN)

C. Data collection

In terms of data gathering, participants provided self-reported responses through a 50-item questionnaire based on the International Personality Item Pool for Big-Five personality traits (Goldberg et al., 2006) and a trust questionnaire developed by Körber (2019). Eye movement data were collected via Tobii Pro Glasses 2, and analyzed with Tobii Pro Lab software, operating at a 50 Hz sampling rate.

The analysis focused on several eye movement metrics divided into three known categories: temporal (e.g., duration and frequency of gaze fixations), spatial (e.g., saccade amplitude and eye movement patterns), and count (e.g., number of fixations and transitions between specific areas of interest) (Boudreau et al., 2009; Lu & Sarter, 2019; Yang et al., 2017). These metrics helped assess the participants' attention distribution and scanning strategies. As illustrated in Figure 28, multiple Areas of Interest (AOIs) were designated on the simulation screens to track

how participants' visual attention varied, especially in response to system malfunctions. The processing and analysis of eye movement data were conducted using the Tobii Pro software suite, IBM SPSS Statistics, and the R programming language.

Figure 28, Areas of Interest (AOI)

D. Design of Experiment

The experiment employed an observational within-subjects design to assess human interaction with autonomous maritime systems. Participants, all exposed to identical conditions, engaged with a simulated mid-size Roll-on/Roll-off (Ro-Ro) vessel autonomously navigating from Horten to Moss in Norway, with additional vessels introduced to simulate realistic maritime traffic (see Figure 29). The simulation tasked participants with monitoring the vessel's autonomous navigation, with the option to manually override the system in response to potential malfunctions. The vessel's navigational control can be selected from three modes of autonomy in the Steering System panel including, Autonomous (NAV), Autopilot (AUTO), and Manual (MANUAL), as shown in Figure 30.

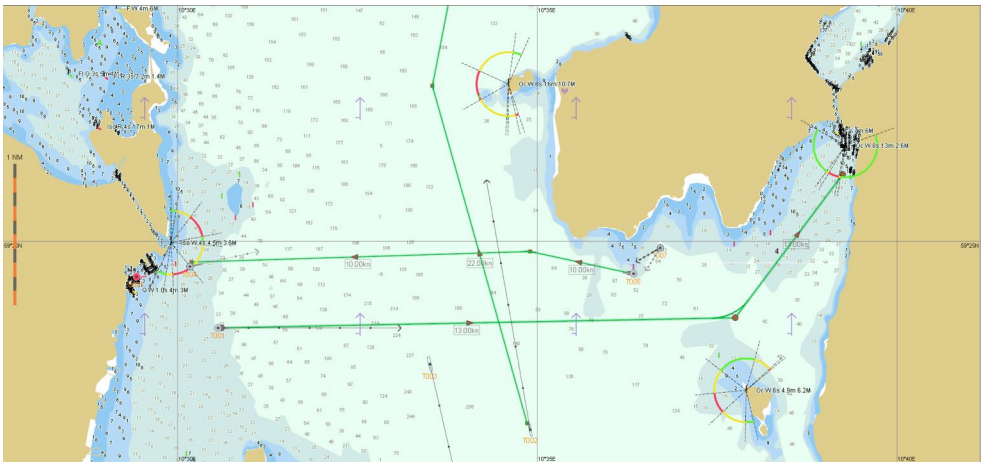


Figure 29, Vessel's traffic environment

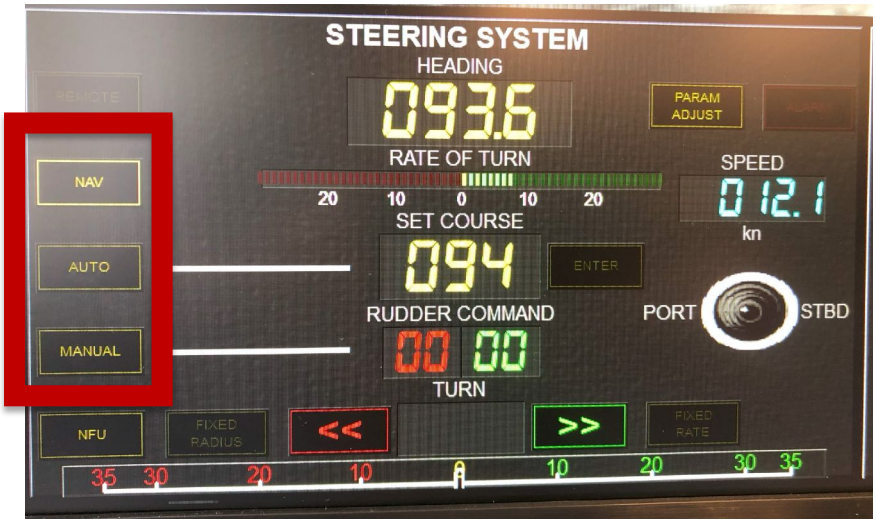


Figure 30, Steering System Panel

E. Procedure

Initially, participants received an overview of the study's objectives, consented to participate, and completed demographic and personality surveys. They were then informed about the study's tools and tasks, including the calibration of an eye-tracking device for precise data gathering.

During the first half of the study, the system functioned without fault, showcasing ideal automation conditions. However, in the latter half, a deliberate malfunction — the halting of a steering gear pump — was introduced to observe participants' responses to unexpected navigational deviations, as shown in ECDIS (Figure 31), and signaled by an alert. Deviation from the course commenced after 10 minutes and lasted for 60 seconds. If the test subject failed to notice and/or take over, the vessel was set to go back on course after the error period. Test subjects were expected to notice the change in course anytime within the 60 seconds. By the end of the study, if any test subject did not notice the deviation, it would be considered a failure to recognize. Trust measurements were recorded before and after this induced error to assess

changes in participants' perceptions of the autonomous system. The entire session lasted about 40 minutes. A graphical representation of this procedure is displayed in Figure 32.

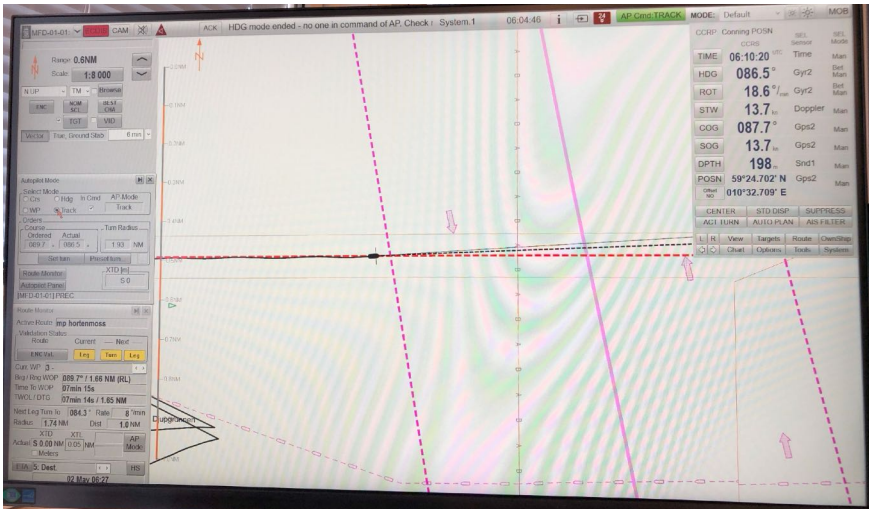


Figure 31, Vessel's deviation from the pre-defined route.

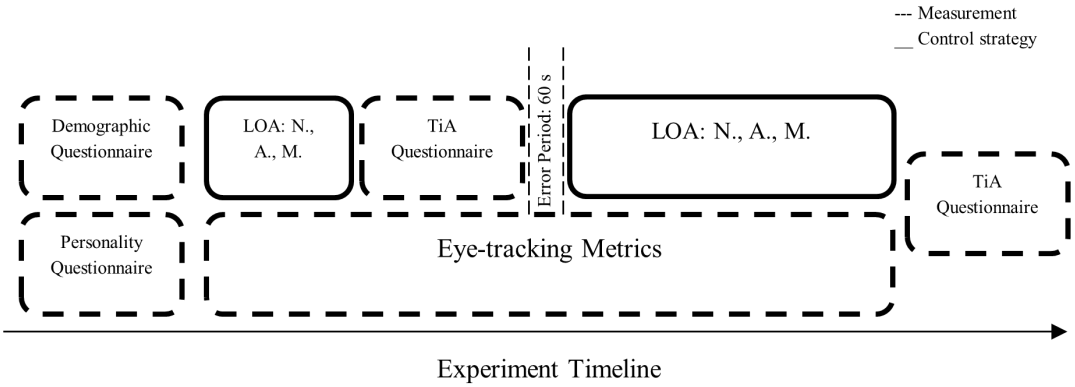


Figure 32, Experiment procedure

6.4.4 Experiment Results

Analysis using a paired samples t-test revealed significant changes in gaze metrics and perceptions of reliability and trust in automation post-error (see Table 7). Specifically, total and average fixation durations increased, indicating heightened attention to Areas of Interest (AOIs) following the error. While spatial metrics such as the average amplitude of saccades showed no change, the total amplitude of saccades increased. Notably, the number of visits, saccades, and fixations to AOIs rose, signifying more intensive engagement with the information provided by the AOIs after the system malfunction. A heatmap of visual attention before and after the error is shown in Figure 33. Perceived reliability and trust in automation both decreased significantly from pre-error to post-error, with reliability showing a larger mean decrease.

Table 7, Results of perceived reliability, trust, and gaze metrics pre- and post-error

	Time 1-Pre-Error (N=28)			Time 2 -Post-Error (N=28)			Paired Differences				
	M	*SD*	*95% CI*	*M*	*SD*	*95% CI*	M_{diff}	*SD*	*95% CI*	*t*	*Cohen's d*
REL	3.46	0.373	[3.32, 3.60]	2.69	0.694	[2.49, 3.01]	0.77	0.616	[0.53, 1.01]	6.651	1.25
TRU	3.68	0.629	[3.40, 3.87]	3.02	1.018	[2.70, 3.46]	0.66	0.82	[0.344, 0.98]	4.279	0.80
Temporal Metrics											
TDF	44963.11	38535.01	[28916, 57216]	101045.39	69226.59	[71718, 122764]	-56082.286	55529.778	[-77614, -34550]	-5.34	1.01
ADF	322.32	69.08	[294.12, 354.42]	390.96	107.84	[346.79, 426.01]	-68.643	108.798	[-110, -26]	-3.33	0.631
Spatial Metrics											
AMS	5.71	0.87	[5.38, 6.03]	5.5	0.90	[5.26, 5.92]	.13964	.59624	[-.09, .37]	1.23	0.234
TAS	4190	1260.82	[3695, 4615]	5162.54	1437.54	[4595, 5639]	-972.53	1267.43	[-1463, -481]	-4.06	0.767
Count Metrics											
NV	56.61	27.87	[45.6, 66.26]	113.04	43.56	[95.19, 127.41]	-56.429	36.839	[-70, -42]	-8.10	1.532
NS	51.54	56.17	[28.10, 69.36]	86.64	56.35	[61.31, 103.62]	-35.107	48.293	[-53, -16]	-3.84	0.727
NF	133.5	102.70	[91.10, 166.5]	253.11	136.54	[195.14, 295.99]	-119.607	111.021	[-162, -76]	-5.70	1.07

Notes. REL = perceived reliability; TRU = trust in automation; TDF = total duration of fixation; ADF = average duration of fixation; AMS = average amplitude of saccades; TAS = total amplitude of saccades; NV = number of visits; NS = number of saccades; NF = number of fixations; M = mean; SD = standard deviation; 95% CI = 95% confidence interval, M_{diff} = mean difference; t = t-statistics

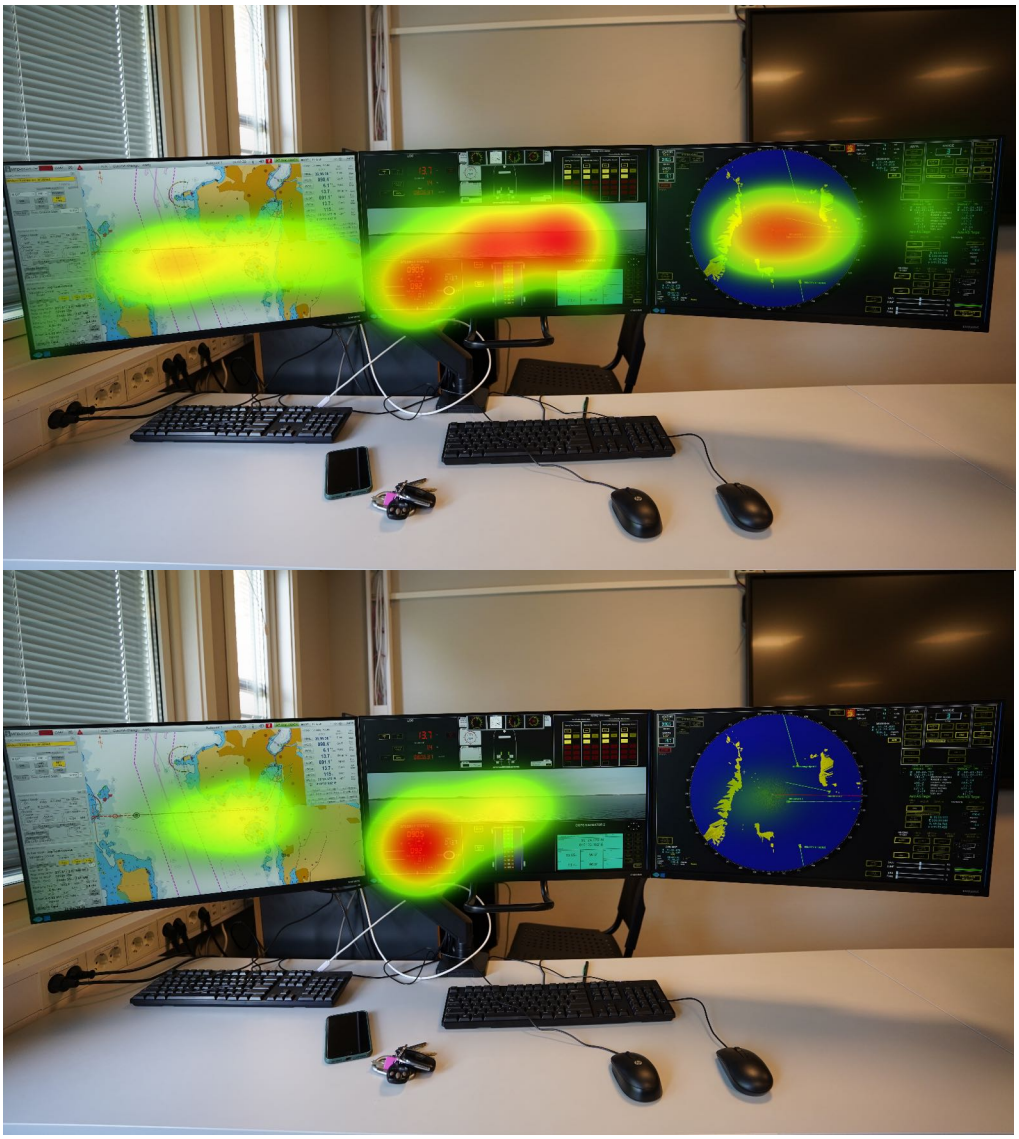


Figure 33, Visual attention prior (top) and post (bottom) system malfunction.

Correlation analysis (Table 8) between perceived reliability, trust in automation, and personality traits revealed a strong positive relationship between reliability and trust at both time points, but no significant correlation with personality traits. Change in trust in automation (dTRU) and the change in perceived reliability (dREL) are also positively correlated, indicating a concurrent decrease in trust and perceived reliability from the two time points. It is important to note that personality traits do not show significant correlations with changes in trust, suggesting that personality may not play a substantial role in the observed decline in trust because of system malfunction.

Table 8, Correlational matrix

Variable	α	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. REL t_1	.40	3.46	0.39										
2. TRU t_1	.77	3.68	0.63	.65** [.37, .82]									
3. REL t_2	.78	2.69	0.68	.44* [.08, .70]	.55** [.22, .77]								
4. TRU t_2	.93	3.02	1.02	.32 [-.06, .62]	.60** [.29, .80]	.79** [.59, .90]							
5. dTRU	_	-0.62	0.82	-.09 [-.45, .29]	-.06 [-.43, .32]	.51** [.18, .74]	.74** [.51, .87]						
6. dREL	_	-0.77	0.62	-.15 [-.49, .24]	.20 [-.19, .53]	.83** [.65, .92]	.66** [.39, .83]	.62** [.33, .81]					
7. E	.82	24.79	6.59	.12 [-.27, .47]	.02 [-.35, .39]	.10 [-.29, .45]	.26 [-.12, .58]	.29 [-.09, .60]	.03 [-.34, .40]				
8. A	.75	28.32	5.89	.12 [-.27, .47]	-.02 [-.39, .36]	-.14 [-.49, .24]	-.11 [-.46, .28]	-.07 [-.43, .31]	-.23 [-.56, .15]	.35 [-.03, .64]			
9. C	.69	25.11	5.82	.08 [-.30, .44]	.35 [-.02, .64]	.13 [-.25, .48]	.14 [-.25, .49]	-.04 [-.41, .34]	.09 [-.29, .45]	-.13 [-.48, .25]	-.02 [-.39, .36]		
10.N	.78	26.86	5.87	.03 [-.34, .40]	.01 [-.36, .38]	.14 [-.25, .49]	.06 [-.32, .43]	.04 [-.34, .41]	.13 [-.25, .48]	-.05 [-.41, .33]	.13 [-.26, .48]	.42* [.05, .69]	
11.O	.73	27.18	5.12	.15 [-.24, .49]	.34 [-.04, .63]	.06 [-.32, .42]	.24 [-.15, .56]	.09 [-.29, .45]	-.03 [-.40, .35]	.01 [-.37, .38]	.24 [-.14, .56]	.37 [-.01, .65]	.04 [-.34, .41]

Note. REL = perceived reliability; TRU = trust in automation; dTRU = change in trust in automation, dREL = change in perceived reliability E = extraversion; A = agreeableness; C = conscientiousness; N= neuroticism; O = openness; t_1 = measurement at time 1, t_2 = measurement at time 2, α = Cronbach's alpha; M = mean; SD = Standard deviation, Values in square brackets indicate the 95% confidence interval for each correlation, * indicates $p < .05$. ** indicates $p < .01$.

Comparing empirical findings with the System Dynamics model showed that perceived reliability and trust decreased by 15.4% and 13.2% respectively, aligning with the model's predictions of a 14% decline in perceived performance and a 12.7% decrease in trust in automation. This congruence between simulated and observed trust dynamics empirically validates the portion of the model's applicability in reflecting real-world trust behavior following system malfunctions.

The proposed system dynamics model illuminates how initial conditions shape trust dynamics in human-automation interactions, confirming the enduring influence of first impressions. Ensuring alignment between perceived performance, trust, and user expectations is vital for autonomous systems' effectiveness. The model suggests that trust can self-perpetuate, provided the system meets expectations, but may diminish with perceived unreliability. Trust recovery after malfunctions depends on the magnitude and timing of the faults, with early issues causing more pronounced and prolonged trust decreases. Empirical findings from the MASS study support the model's hypotheses, showing that personality traits do not significantly affect trust variations due to system malfunctions. The model's structural focus on trust dynamics, rather than specific contextual details, offers broad applicability across scenarios but also necessitates adjustments for precise contextual representation. Overall, the study offers a flexible, and adaptable system dynamics model that can generate testable hypotheses about trust evolution,

offering a tool for future research to investigate trust across various contexts and automation levels.

7 Synthesis of the Results and General Reflection

This chapter consolidates the principal outcomes of this dissertation, offering an overarching reflection on the entire research journey. The connections between each article and their central insights are depicted in Figure 34. Additionally, this chapter outlines the research's limitations and suggests directions for future inquiry in the field of Human-Automation Interaction (HAI).

Article 1 (Poornikoo & Øvergård, 2023) initiated an in-depth examination into the realm of theory development and model building, centering specifically on models of trust in automation (TiA) and the application of scientific criteria for evaluating these models. The objective of this article was to explore the foundational aspects of HFE models and to assess whether current modeling endeavors offer a reliable and comprehensive understanding of trust in automation in today's complex sociotechnical systems. It revealed a prevalent reliance on conceptual (variance) modeling as the primary approach for depicting variations in trust. This approach has been instrumental in identifying potential factors influencing trust and has enabled further analysis through empirical research. However, despite the assertion that trust is inherently dynamic and subject to change over time, the predominant modeling language and focus on static causal factors, rather than on interrelationship dynamics, renders these models somewhat restrictive. As a result, studies on TiA and cognitive psychology continue to depend heavily on static models and traditional experimental methodologies, often defaulting to statistical data analysis. This overreliance poses a critical issue: the statistical framework utilized in the analysis inadvertently becomes a replacement for an actual static model. This approach, primarily designed to structure data and hypothesize relationships in the absence of concrete understanding, falls short of accurately depicting social processes and dynamics. Notably, standard methods such as analysis of variance and regression are especially inadequate for capturing complex social interactions, with time-based dynamics being particularly underrepresented (Levine, 2000; Levine & Doyle, 2002; Levine & Fitzgerald, 1992).
Several other issues were discussed in Article 1 regarding the efficacy of modeling approaches in relation to satisfying the proposed scientific criteria. Specifically, the conventional variance conceptual models do not allow for scientific checks of the entire model, as this requires the operationalization of all variables and the relationship between them. As a result, the empirical adequacy of the models becomes somewhat limited. The choice of variance modeling approach is mainly in line with the reductionist view of science, despite the ongoing debates regarding viewing the HAI and HFE as a systems discipline. Article 1 further examined the computational models of trust in automation, revealing that such models are better in testability/falsifiability criterion and predictive power. Nevertheless, they heavily rely on data and consequently, they become data models with limited applicability and generalizability to be adapted in new settings. The suggestions in Article 1 pointed towards adopting novel modeling approaches that can better satisfy the proposed criteria.

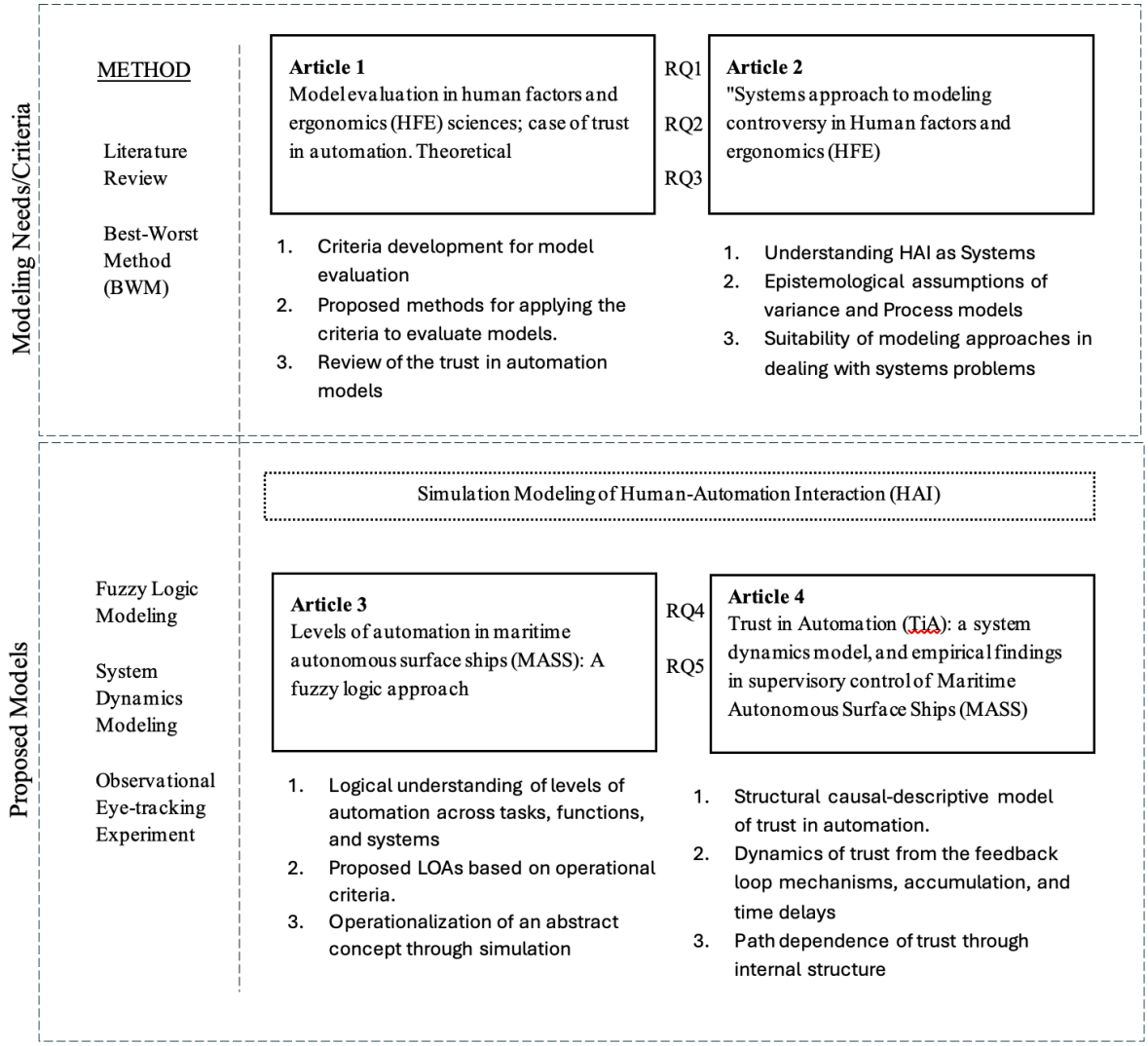


Figure 34, Overview of the four articles, key insights, and the methods

In Article 2 (Poornikoo & Mansouri, 2023), assumptions of the variance models were examined to realize the notion of cause and effect in such models. The primary definition of cause and effect in the variance model implies that the cause is necessary and sufficient for the outcome, and the outcome inevitably occurs when necessary and sufficient causes are available. This overly simplified static assumption can only be applied for a fraction of the time in which the phenomenon is being studied. Yet, the entire broader picture of continuous human interaction with the automation and dynamic environment cannot be studied with such assumptions. On the other hand, process models assume that causation consists of necessary conditions in sequence, and the outcome may not occur even when the causal conditions are present. The logical differences in the two modeling paradigms can be explained as: if X, then Y; if more X then more Y (in variance models), whereas if not X, then not Y (in process models) (Markus & Robey, 1988; Payne et al., 2017; Poole et al., 2000).

As discussed in Article 2, while the assumptions underlying variance models may be sufficient for comparisons among entities and general linear models, they are rather ill-suited for studying social entities that are prone to nonlinear dynamic interactions. On the other hand, assumptions

of process (casual-descriptive) models, can account for the reciprocal causality, time-lagged, and historical development of trust in automation.

The primary objective of Articles 1 and 2 was to map out the current modeling practices within Human Factors and Ergonomics (HFE) and to delineate the crucial scientific criteria necessary for the evaluation of models within this domain. The findings from both articles underscore that simulation modeling emerges as a particularly promising strategy for addressing the multifaceted nature of Human-Automation Interaction (HAI) challenges.

With this in mind, Article 3 (Poornikoo & Øvergård, 2022) focused on levels of automation, a critical aspect of human-automation interaction. It was initiated by reviewing the existing taxonomies and levels and defining the concept of automation across systems, functions, and tasks. An autonomous system in this regard was defined as a system capable of performing functions autonomously (Poornikoo & Øvergård, 2022), such as route planning and autonomous navigation. A fuzzy logic simulation model was crafted to describe levels of automation according to operational criteria, aiming to establish a common language among scholars and practitioners engaged in Maritime Autonomous Surface Ships (MASS). This model aspires to integrate Levels of Automation (LOAs) effectively into both research and modeling endeavors.

Article 4 explores a different simulation technique to analyze trust in automation, emphasizing the internal structure and feedback loops to predict system behavior. Unlike previous conceptual models, this study narrows its focus to the feedback mechanisms, revealing the universal nature of trust's path dependence in automation contexts. It postulates that individual variances significantly affect the rates at which trust adjusts in response to automation performance. A critical insight from this research is the model's sensitivity to system malfunctions; the timing and magnitude of errors significantly influence trust levels, with early-stage errors causing more substantial damage. Indeed, this hypothesis necessitates further empirical support. Article 4 also undertakes empirical tests to examine the effects of errors on trust and perceived reliability and explores how such incidents alter monitoring behavior. Experimental results not only support the model's predictions but also highlight the impact of trust on supervisory control within Maritime Autonomous Surface Ships (MASS). The findings indicate that trust levels significantly dictate the monitoring frequency and attention distribution across different Areas of Interest (AOIs), suggesting that inadequate trust can lead to a disproportionate focus on specific AOIs at the expense of others, thus underscoring the critical role of trust in the effective supervision of autonomous systems.

7.1 General Reflection

Hollnagel (2002) reminds us about the practices of modeling in the study of human-machine systems. The first practice focuses on the *how* of modeling, emphasizing the structure and content of the models. This approach assumes the necessity of modeling and aims at devising the most effective or sophisticated methods to fulfill this need. It is primarily interested in the architecture of the model, including its components and their configuration. The second practice concentrates on *what* is being modeled, i.e., the purpose and outcome of the modeling, prioritizing the model's functionality or performance over its structure. This functional approach values what the model aims to replicate or predict rather than the intricacies of its construction. This method is more concerned with the end function and results, choosing the

most apt solution based on the phenomena being addressed rather than a fixed, bottom-up construction of specific model components. He further emphasizes *time* and *control* – indispensable aspects of human action – as neglected issues in cognitive ergonomics models.
It can be maintained that the structure and function within modeling are interconnected and mutually reinforcing concepts. An in-depth understanding of the phenomenon being modeled, including its intrinsic characteristics, is essential before a meaningful and explanatory model can be constructed. In the domain of cognitive ergonomics, particularly in Human-Automation Interaction (HAI), it appears that fundamental concepts such as *change*, *time*, and *state* are often overlooked in contemporary ergonomic models and methodologies. To further elaborate, several decades ago, Lewin (1951) theorized the field theory, that psychological phenomena take place within a field termed a "life space" or "psychological field", characterized by the dynamics of interconnections and positionalities. Behavior is viewed as a product of the interaction between an individual and their life space, summarized in $B = f(LS)$. Life space (LS) encompasses both the individual (P) and their surrounding environment (E), leading to the revised equation $B = f(P, E)$. This formula suggests that behavior emerges from the interplay between a person and their environment, meaning that behavior is represented as *locomotion* from one region of the life space to another, regions that attract or repel having valence. Lewin further argues that any behavior or changes in behavior depends only on the psychological field *at that time*. Field theory asserts that behavioral change $\frac{dx}{dt}$ at the time t depends on the situation S^t at the time t, or $\frac{dx}{dt} = F(S^t)$, and in closed systems, also depends on a past situation S^{t-n}. This somewhat overlooked principle offers valuable insights for today's modeling practices. Specifically, recognizing that phenomena are dynamic suggests that their characteristics can evolve over time. Consequently, factors that were crucial at one stage may become less influential later, leading to variations in the strength of correlations as time progresses. This phenomenon reflects the challenges associated with nonlinear dynamic complexity in modeling (Guastello, 2017).

Modeling human behavior in human-automation systems primarily involves cognitive processes and dynamic evolution. In scientific research, models of human behavior serve not only to unify diverse empirical findings but also to offer a systematic understanding of the mechanisms underlying human behavior (Gauch, 2012). A cognitive model then must be designed to reflect such mechanisms with adequate intricacy. While verbal descriptions or conceptual models are crucial in guiding many experimental studies for identifying causal factors, they often fall short of explaining the kind of relationships among variables and the dynamics of human behavior in making predictions. This limitation becomes particularly apparent given the complexity of the human cognitive and motor systems, which verbal descriptions or a simple box-and-arrow representation may struggle to quantify effectively.

Computer-assisted modeling (i.e., simulation) can provide a more detailed and quantifiable method of representing the intricate relationships within human-automation systems. These models are not only useful for solving the limitations of verbal descriptions but also play a crucial role in guiding researchers in their data collection efforts. Furthermore, simulation models offer a great opportunity to formulate testable hypotheses in dynamic scenarios.

Programming a computer to imitate human actions represents a more thorough level of understanding than is typically achieved through conventional box-and-arrow diagrams. Successful cognitive models, when based on certain assumptions, can produce human-like behaviors. By altering these assumptions, researchers can observe changes in the model's behavior. Such explorations provide a basis for designing experimental conditions likely to yield measurable effects, further enhancing the understanding of human behavior.

7.2 Research Limitations

The objective of this doctoral dissertation was to create a comprehensive and robust body of work. However, it is important to acknowledge certain limitations that were encountered. These limitations were primarily due to the time constraints associated with this academic endeavor and the theoretical nature of the research undertaken. A detailed discussion of these constraints and their implications is provided in the subsequent sections.

7.2.1 Model Evaluation

Article 1 introduced a framework for assessing Human Factors and Ergonomics (HFE) constructs, setting the stage for assessing models of trust in automation. Initially, the proposed evaluation criteria derive from established scientific criteria; however, there is room for adaptation to tailor these criteria more closely to specific HFE models in subsequent research efforts. Additionally, the prioritization of these criteria through the Best-Worst Method (BWM) is inherently influenced by the subjective perspectives of the researchers involved. Fostering agreement among researchers regarding the evaluation of models could lead to more consistent and reliable assessments. Moreover, the research recognizes that a one-size-fits-all approach may not be applicable, and aligning the evaluation criteria with the application context could facilitate more precise and relevant model selection. This approach would enable practitioners to choose the most suitable model based on the specific requirements of their situation, whether it be for theoretical exploration or practical design purposes.

7.2.2 Modeling Epistemology

Article 2 explored two distinct modeling approaches (i.e., variance, and process) with regards to their epistemological assumptions and utility in HFE modeling efforts. While the paper discusses the limitations of current HFE models in adopting a variance approach, it primarily focuses on outcome-oriented and mathematical simulation models. That means that the exploration of alternative modeling approaches is not deeply explored. The article provides a theoretical discussion on the need for systems approaches in HFE modeling but lacks studies or case examples to support the proposed benefits of integrating variance and process models.

7.2.3 Level of Automation (LOA) Fuzzy Logic Model

The fuzzy model of LOAs encompasses several limitations that warrant further investigation. One of the primary constraints lies in the subjective selection of membership functions within the fuzzy logic model, which can significantly vary based on the specific scenario being analyzed. This subjectivity calls for additional studies aimed at identifying the most appropriate

membership functions tailored to distinct Levels of Automation (LOAs). An advanced strategy, such as employing a weighted fuzzy approach, is suggested to refine this process. Moreover, the conceptual foundation of the proposed model is built upon existing presumptions related to human-automation interaction (HAI) dynamics and the operational definitions of LOAs. These underlying assumptions form the backbone of the current framework but may not encompass all possible variations and complexities inherent in HAI scenarios. Consequently, future studies should delve deeper into these presumptions, examining their validity and applicability in the context of fuzzy logic models for LOAs, thereby enhancing the model's relevance and accuracy. Lastly, the practical implications and efficiency of the proposed approach, particularly in the context of Maritime Autonomous Surface Ships (MASS), have yet to be fully explored. Empirical testing and real-world application of the model within this domain are critical to ascertain its efficacy and identify potential areas for refinement. Adjustments may be required to align the model more closely with the unique challenges and requirements of MASS operations.

7.2.4 Trust in Automation (TiA) System Dynamics Model

This study has introduced a system dynamics model centered on trust in automation, illustrating the potential of simulation modeling to generate dynamic, testable hypotheses regarding trust and related behaviors. The model stands out due to its dynamic nature as opposed to static conceptual frameworks, offering a versatile tool that adapts to empirical data. This characteristic enables the examination of various theories concerning the development and evolution of trust, enhancing its utility for future research in different settings by altering parameters or model structures. Such flexibility is vital for accurately representing complex interactions between humans and automated systems in diverse operational contexts. However, the model predominantly emphasizes the structural dynamics of trust, sidelining the contextual specifics that might influence trust dynamics in particular scenarios. This lack of specific situational details could limit the model’s immediate applicability, requiring modifications to suit distinct environments and conditions for precise simulation outcomes.

For system designers, the model provides insights into integrating psychological and behavioral dimensions of trust from the early stages of development, aiming for systems that users can trust and understand comprehensively. However, the model's current iteration specifically addresses scenarios with high levels of automation, such as fully autonomous driving or autonomous maritime navigation, focusing mainly on situations where human operators supervise and intervene when necessary. This choice limits the model's exploration of trust dynamics across varying levels of automation and does not consider the transitions between different automation levels or how trust fluctuates within these transitions. To enhance the model's relevance and applicability, future versions could introduce modular components to accommodate different operational contexts, system reliability levels, user experiences, and environmental conditions. Such modifications would extend the model's applicability, making it a more practical tool for diverse applications.

8 Conclusion

This chapter concludes this dissertation's research outcomes and scholarly contributions, aligning with the research questions established in the introduction chapter. A summary of the outcomes from each study, along with their theoretical and practical contributions is provided in Table 9.

The advent of Maritime Autonomous Surface Ships (MASS) can represent a significant leap forward in the maritime industry, promising to redefine sea transportation's efficiency, safety, and economics. However, this technological advance brings forward the complex interplay between human operators and autonomous systems, particularly in the context of Shore Control Centers (SCCs), where remote operators play critical roles. The success of integrating MASS into the global shipping infrastructure depends not just on technological advancements but equally on understanding and optimizing Human-Automation Interaction (HAI). The transition to supervisory control roles introduces a paradigm shift in operational dynamics. Remote operators are tasked with maintaining oversight over multiple vessels simultaneously, each possibly facing different sea conditions and operational challenges. This multi-vessel management can significantly amplify the cognitive load, requiring operators to prioritize information effectively and make swift decisions to ensure safety and efficiency. One of the primary concerns is the risk of over-reliance on automation, which may lead to complacency and reduced situational awareness (Endsley, 1996). The remote nature of operation may exacerbate these issues, as operators are removed from the immediate physical environment of the vessels they control. Moreover, the unpredictable and dynamic nature of maritime environments makes complete autonomy a challenging goal; remote operators must be prepared to take control in complex or emergency situations.

To address these challenges and leverage the full potential of MASS, it is imperative to develop scientific and robust models of HAI. These models should account for the unique demands of maritime environments and the specific roles of remote operators. By understanding the cognitive, psychological, and social factors that influence remote operators' performance, researchers and practitioners can design more intuitive and effective interfaces and decision-support systems. Effective HAI models can guide the development of training programs tailored to the needs of remote operators, focusing on critical skills such as situational awareness, decision-making under uncertainty, and effective communication with autonomous systems (Bachari-Lafteh & Harati-Mokhtari, 2021; Emad & Ghosh, 2023; Wright, 2020). Moreover, these models can help identify potential sources of error, the operators' responses, and cognitive overload, enabling the design of systems that support operators' decision-making processes and reduce the likelihood of accidents. Two pivotal aspects of these models are the Levels of Automation (LOAs) and Trust in Automation (TiA). Understanding and accurately modeling these dimensions are crucial for designing systems that effectively balance human supervisory control of autonomous capabilities.

The concept of LOAs in HAI models refers to the range of functions and decision-making capabilities allocated between humans and machines. These levels span from full human control to full automation, with various intermediate stages where tasks are shared or divided differently between humans and systems. Properly modeling LOAs involves identifying the optimal distribution of tasks that maximizes system performance and human well-being. In maritime autonomous systems, such as Maritime Autonomous Surface Ships (MASS), defining LOAs is critical due to the complex, dynamic nature of maritime environments and the high stakes involved in navigation and operations. A robust LOA model for HAI in maritime contexts should provide a framework for dynamically adjusting the level of automation based on specific conditions, operator workload, and system reliability. Such models help in designing interfaces and control systems that allow for seamless transitions between levels of automation, minimizing potential confusion and ensuring that operators remain engaged and prepared to take control when necessary.

Trust in automation is a multifaceted concept that plays a crucial role in HAI. It influences how much reliance operators place on automated systems and how they interact with them. Inappropriate levels of trust—either too high (overreliance) or too low (underutilization)—can lead to suboptimal system performance and increase the risk of accidents (Lee & See, 2004; Parasuraman & Riley, 1997). Modeling TiA requires a deep understanding of the factors that influence trust, such as the system's reliability, transparency, predictability, and the operator's personal experiences and biases. Nonetheless, it also requires an understanding of the relationships among these factors in a closed-loop dynamic HAI process. A well-constructed TiA model enables the design of automated systems that communicate their intentions and limitations clearly, fostering appropriate levels of trust. Such systems should provide operators with feedback and explanations for their actions, especially when autonomous decisions deviate from the operator's expectations. By ensuring that operators understand the capabilities and reasoning of the automated systems, these models can enhance collaboration between humans and machines, leading to more effective decision-making and improved safety outcomes.

In response to the growing scrutiny regarding the validity of Human Factors and Ergonomics (HFE) models, as well as the need for flexible yet credible HAI models, this dissertation concentrated on the importance of models and modeling within Human-Automation Interaction (HAI), particularly emphasizing Trust in Automation (TiA) and Levels of Automation (LOA) as central themes for modeling exploration. This dissertation commenced by exploring the significance of scientific modeling and developed criteria that can be utilized to assess the relative scientific credibility of various models. Furthermore, models of Trust in Automation (TiA) were assessed against these criteria not only to showcase the use of the criteria but also to understand the TiA modeling efforts in the literature. On the other hand, epistemological accounts of modeling efforts were investigated, to realize the suitability of each approach for modeling HAI. The findings suggested simulation as a viable approach to tackle the complexities in modeling TiA and LOA within the context of HAI and supervisory control of MASS. By incorporating models of Trust in Automation (TiA) and Levels of Automation (LOA), simulation offers a powerful tool for examining complex interactions and dynamics that are difficult, if not impossible, to study in real-world settings due to safety, cost, and

practicality concerns. Through simulation, researchers can dissect the underlying mechanisms of TiA and LOA, providing a deeper understanding of how and why certain interactions affect human performance and system efficiency. For example, simulations can reveal how changes in the performance of autonomous systems impact operators' trust levels and their ability to make informed decisions. Another significant advantage of simulation modeling is its ability to predict outcomes of various scenarios. This predictive capability is invaluable for informing the development of MASS, allowing designers to anticipate human factors challenges and address them proactively.

While specific simulations may be designed around particular scenarios or types of MASS operations, the principles and findings derived from this dissertation may have broader applicability. By identifying general patterns in how humans interact with automation, simulation modeling contributes to a body of knowledge that can inform the design and operation of a wide range of automated systems, beyond the maritime context. This generalizability makes simulation an invaluable tool in the broader field of HAI research.

Table 9, Summary of key findings and contributions of this dissertation

RQs	Article	Key findings	Contributions
RQ1: What constitutes the essential criteria for evaluating models within the domain of Human Factors and Ergonomics (HFE) research?	**Article 1** Poornikoo, M., & Øvergård, K. I. (2023). Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation. Theoretical Issues in Ergonomics Science, 1-37. https://doi.org/10.1080/1463922X.2023.2233591	Seven scientific criteria were developed including: Falsifiability/testability, Predictive power, Explanatory power, Empirical adequacy, Pragmatic adequacy, Human as an active agent, and Dynamic properties. These criteria were applied to Trust in Automation models using the Best-Worst Method (BWM).	The study offers a set of checklists for evaluating the scientific credibility of Human Factors and Ergonomics (HFE) models. These criteria can be used as reference tools for comparing different models. The applicability of findings can go beyond TiA models and can be applied to other cognitive models such as Situation Awareness and Mental Workload.
RQ2: What is the current state of Trust in Automation (TiA) models according to the criteria in RQ1?	**Article 1** Poornikoo, M., & Øvergård, K. I. (2023). Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation. Theoretical Issues in Ergonomics Science, 1-37. https://doi.org/10.1080/1463922X.2023.2233591	Two clusters of TiA models were identified in the literature, conceptual and computational models. It was revealed that conceptual models are better at incorporating causal factors and have explanatory power. Contrarily, computational models have the advantage of higher testability scores but are limited in generalizability and explanatory power.	The findings of this study highlighted the advantages and disadvantages of different modeling approaches in the conceptualization and prediction of TiA. This study contributed to the theoretical understanding of TiA research *programme*, and the progression of the field.
RQ3: Are the epistemological assumptions in different modeling approaches appropriate for studying human-automation interactions?	**Article 2** Poornikoo M., Mansouri M. (2023), Systems approach to modeling controversy in Human factors and ergonomics (HFE), 18th Annual System of Systems Engineering Conference (SoSe), Lille, France, 2023, pp. 1-8, https://doi.org/10.1109/SoSE59841.2023.10178634	This study distinguished static (variance) and dynamic (process) modeling approaches and discussed the underlying epistemological assumptions in defining causal and effect relationships.	The findings and discussions highlighted the shortcomings of the assumptions of variance models in modeling human agents and behavior. The study contributed to the theoretical and epistemological assumptions of various modeling approaches and recommended utilizing process (dynamic) modeling practices in studying human performance in sociotechnical systems.
RQ4: How to effectively model Levels of Automation (LOAs) for Maritime	**Article 3** Poornikoo, M., & Øvergård, K. I. (2022). Levels of automation in maritime autonomous surface ships (MASS): A fuzzy logic approach. Maritime Economics & Logistics, 24(2), 278-301.	This study reviewed the existing levels of automation and highlighted their ambiguity and limitations in operational scenarios. Utilizing the LOA's basic constituents and fuzzy logic approach, the study	This study is the first attempt to quantify and operationalize the levels of automation for broader use and applicability. The main contribution of this model lies in using an approach that can take human judgements as inputs and

Autonomous Surface Ship (MASS)?	https://doi.org/10.1057/s41278-022-00215-z	developed a simulation model for levels of automation.	return the corresponding LOA, offering a universal language for defining the LOAs not as discrete taxonomies but a continuous spectrum.
RQ5: How can Trust in Automation (TiA) be dynamically modeled based on its internal structures?	**Article 4** Poornikoo M., Gyldensten W., Vesin B., Øvergård, K. I. (Submitted) Trust in Automation (TiA): a system dynamics model, and empirical findings in supervisory control of Maritime Autonomous Surface Ships (MASS), Human-Computer Interaction	Building on the findings from RQ1 and RQ2, this study constructs a structural causal model of trust in automation that can portray trust dynamics in three stages: trust formation, trust loss, and trust repair. Using system dynamics simulation model, the model generates behavior based on its internal structure. The study was further examined through an empirical study for validation.	The contribution of this study includes: (1) developing a dynamic model with few assumptions that can be tested and refuted. (2) incorporating individual variability as part of the adjustment time in response to observed performance. (3) proposing several scenario-based behaviors such as the effect of error time on trust, or initial trust/expectation mismatch.

8.1 Future Research Recommendations

The findings from this dissertation mark a significant stride towards enriching future research in the fields of Human-Automation Interaction (HAI) and cognitive ergonomics. The simulation models introduced in this dissertation represent preliminary efforts to explore these complex domains with a new perspective. While these initial models might appear rudimentary to simulation experts, including those specializing in system dynamics and fuzzy logic, they lay a foundational groundwork for the development of more intricate and thorough models in subsequent research endeavors. Moreover, while the current system dynamics model of trust in automation is tailored to scenarios involving high levels of automation, such as fully autonomous systems, there is a clear pathway for further refinement. Specifically, there is an opportunity to enhance the integration of Trust in Automation (TiA) system dynamics with various levels of automation articulated through fuzzy logic.

Regarding model evaluation, the framework introduced in this dissertation holds potential for broader application, extending beyond trust models to evaluate other critical HFE constructs such as situation awareness and mental workload. Also, the proposed system dynamics TiA model itself presents an opportunity for rigorous assessment using these criteria in subsequent studies.

9 References

Abilio Ramos, M., Utne, I. B., & Mosleh, A. (2019). Collision avoidance on maritime autonomous surface ships: Operators' tasks and human failure events. *Safety Science*, *116*, 33–44. https://doi.org/10.1016/j.ssci.2019.02.038

Adams, B. D., Bruyn, L. E., & Houde, S. (2003). *Trust in Automated Systems, Literature Review*. Humansystems Incorporated.

Ahvenjärvi, S. (2016). The Human Element and Autonomous Ships. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, *10*(3), 517–521. https://doi.org/10.12716/1001.10.03.18

Aiello, G., Giallanza, A., & Mascarella, G. (2020). Towards Shipping 4.0. A preliminary gap analysis. *Procedia Manufacturing*, *42*, 24–29.

Akash, K., Reid, T., & Jain, N. (2018). Adaptive Probabilistic Classification of Dynamic Processes: A Case Study on Human Trust in Automation. *2018 Annual American Control Conference (ACC)*, 246–251. https://doi.org/10.23919/ACC.2018.8431132

Akash, K., Wan-Lin Hu, Reid, T., & Jain, N. (2017). Dynamic modeling of trust in human-machine interactions. *2017 American Control Conference (ACC)*, 1542–1548. https://doi.org/10.23919/ACC.2017.7963172

Albus, J., & Antsaklis, P. J. (1998). Panel discussion: Autonomy in engineering systems: What is it and why is it important? Setting the stage: Some autonomous thoughts on autonomy. *Proceedings of the 1998 IEEE International Symposium on Intelligent Control (ISIC) Held Jointly with IEEE International Symposium on Computational Intelligence in Robotics and Automation (CIRA) Intell*, 520–521. https://ieeexplore.ieee.org/abstract/document/713716/

Aldrich, H. E. (2001). Who wants to be an evolutionary theorist? Remarks on the occasion of the year 2000 OMT distinguished scholarly career award presentation. *Journal of Management Inquiry*, *10*(2), 115–127.

Alharahsheh, H. H., & Pius, A. (2020). A review of key paradigms: Positivism VS interpretivism. *Global Academic Journal of Humanities and Social Sciences*, *2*(3), 39–43.

Allen, P. M., & Varga, L. (2007). Complexity: The co-evolution of epistemology, axiology and ontology. *Nonlinear Dynamics, Psychology, and Life Sciences*, *11*(1), 19.

Alonso, V., & De La Puente, P. (2018). System transparency in shared autonomy: A mini review. *Frontiers in Neurorobotics*, *12*, 83.

Atoyan, H., Duquet, J.-R., & Robert, J.-M. (2006). Trust in new decision aid systems. *Proceedings of the 18th International Conference on Association Francophone d'Interaction Homme-Machine - IHM '06*, 115–122. https://doi.org/10.1145/1132736.1132751

Azevedo-Sa, H., Jayaraman, S. K., Esterwood, C. T., Yang, X. J., Robert, L. P., & Tilbury, D. M. (2021). Real-Time Estimation of Drivers' Trust in Automated Driving Systems. *International Journal of Social Robotics*, *13*(8), 1911–1927. https://doi.org/10.1007/s12369-020-00694-1

Bacharach, S. B. (1989). Organizational theories: Some criteria for evaluation. *Academy of Management Review*, *14*(4), 496–515.

Bachari-Lafteh, M., & Harati-Mokhtari, A. (2021). Operator's skills and knowledge requirement in autonomous ships control centre. *Journal of International Maritime Safety, Environmental Affairs, and Shipping*, *5*(2), 74–83. https://doi.org/10.1080/25725084.2021.1949842

Bacon, M. (2012). *Pragmatism: An introduction*. Polity. https://books.google.com/books?hl=no&lr=&id=XtX-O7wbn4EC&oi=fnd&pg=PR5&dq=Pragmatism&ots=fb-_y6cFPP&sig=3U6K7jGBGxUGX49xA9EkepBJ3qs

Bagheri, N., & Jamieson, G. A. (2004). The impact of context-related reliability on automation failure detection and scanning behaviour. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, *1*, 212–217. https://ieeexplore.ieee.org/abstract/document/1398299/

Bainbridge, L. (1983). Ironies of automation. In *Analysis, design and evaluation of man–machine systems* (pp. 129–135). Elsevier.

Bala, B. K., Arshad, F. M., & Noh, K. M. (2017). *System Dynamics*. Springer Singapore. https://doi.org/10.1007/978-981-10-2045-2

Barber, B. (1983). *The logic and limits of trust*. Rutgers University Press.

Barlas, Y. (1996). Formal aspects of model validity and validation in system dynamics. *System Dynamics Review*, *12*(3), 183–210. https://doi.org/10.1002/(SICI)1099-1727(199623)12:3<183::AID-SDR103>3.0.CO;2-4

Barlas, Y. (2018). Credibility, Validity and Testing of Dynamic Simulation Models. In M. S. Obaidat, T. Ören, & Y. Merkuryev (Eds.), *Simulation and Modeling Methodologies, Technologies and Applications* (Vol. 676, pp. 3–15). Springer International Publishing. https://doi.org/10.1007/978-3-319-69832-8_1

Barlas, Y., & Carpenter, S. (1990). Philosophical roots of model validation: Two paradigms. *System Dynamics Review*, *6*(2), 148–166.

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction*, *3*(2), 74. https://doi.org/10.5898/JHRI.3.2.Beer

Beggiato, M., & Krems, J. F. (2013). The evolution of mental model, trust and acceptance of adaptive cruise control in relation to initial information. *Transportation Research Part F: Traffic Psychology and Behaviour*, *18*, 47–57.

Beisbart, C. (2012). How can computer simulations produce new knowledge? *European Journal for Philosophy of Science*, *2*(3), 395–434. https://doi.org/10.1007/s13194-012-0049-7

Bhakthavatsalam, S., & Cartwright, N. (2017). What's so special about empirical adequacy? *European Journal for Philosophy of Science*, *7*(3), 445–465. https://doi.org/10.1007/s13194-017-0171-7

Bhaskar, R. (2013). *A realist theory of science*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9780203090732/realist-theory-science-roy-bhaskar

Bindewald, J. M., Rusnock, C. F., & Miller, M. E. (2018). Measuring Human Trust Behavior in Human-Machine Teams. In D. N. Cassenti (Ed.), *Advances in Human Factors in Simulation and Modeling* (Vol. 591, pp. 47–58). Springer International Publishing. https://doi.org/10.1007/978-3-319-60591-3_5

Biros, D. P., Daly, M., & Gunsch, G. (2004). The influence of task load and automation trust on deception detection. *Group Decision and Negotiation*, *13*(2), 173–189.

Blalock, H. M. (1969). *Theory construction: From verbal to mathematical formulations*. Prentice-Hall Englewood Cliffs, NJ.

Bossel, H. (2007). *Systems and Models: Complexity, Dynamics, Evolution, Sustainability*. BoD – Books on Demand.

Boubin, J. G., Rusnock, C. F., & Bindewald, J. M. (2017). Quantifying Compliance and Reliance Trust Behaviors to Influence Trust in Human-Automation Teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 750–754. https://doi.org/10.1177/1541931213601672

Boudreau, C., McCubbins, M. D., & Coulson, S. (2009). Knowing when to trust others: An ERP study of decision making after receiving information from unknown people. *Social Cognitive and Affective Neuroscience*, *4*(1), 23–34.

Bratley, P., Fox, B. L., & Schrage, L. E. (2011). *A guide to simulation*. Springer Science & Business Media. https://www.google.com/books?hl=no&lr=&id=XHnkBwAAQBAJ&oi=fnd&pg=PR18&dq=Bratley+et+al.,+simulation&ots=fI8pvE8-Gf&sig=U6pvofAkMKAunBUErc0espk_TS4

Bridger, R. S. (n.d.). *Human Factors Integration in the Systems Lifecycle*. Retrieved March 14, 2024, from https://www.researchgate.net/profile/Robert-Bridger/publication/269093944_Keynote_Address_Human_Factors_Integration_in_the_Systems_Lifecycle/links/547f23e60cf2d2200edeb99d/Keynote-Address-Human-Factors-Integration-in-the-Systems-Lifecycle.pdf

Bruner, E. M. (1986). *The anthropology of experience*. University of Illinois Press.

Burmeister, H.-C., Bruhn, W., Rødseth, Ø. J., & Porathe, T. (2014). Autonomous Unmanned Merchant Vessel and its Contribution towards the e-Navigation Implementation: The MUNIN Perspective. *International Journal of E-Navigation and Maritime Economy*, *1*, 1–13. https://doi.org/10.1016/j.enavi.2014.12.002

Caple, D. (2008). Emerging challenges to the ergonomics domain. *Ergonomics*, *51*(1), 49–54.

Card, S. K., Moran, T. P., & Newell, A. (2005). The model human processor. *Ergonomics: Major Writings*, *2*, 382.

Card, S., Moran, T., & Newell, A. (1983). *The Psychology of Human Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Carson, E. R., & Flood, R. L. (1990). Model validation: Philosophy, methodology and examples. *Transactions of the Institute of Measurement and Control*, *12*(4), 178–185. https://doi.org/10.1177/014233129001200404

Cartwright, N., & Montuschi, E. (Eds.). (2014). *Philosophy of social science: A new introduction* (First edition). Oxford University Press.

Cass, E. M. (2011). *Can situation awareness be predicted?: Investigating relationships between CogScreen-AE and pilot situation awareness* [PhD Thesis]. Carleton University.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model* (Vol. 18). John Wiley & Sons.

Chapanis, A. (1988). Some Generalizations about Generalization. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *30*(3), 253–267. https://doi.org/10.1177/001872088803000301

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2018). Planning with trust for human-robot collaboration. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction*, 307–315.

Chen, M., Nikolaidis, S., Soh, H., Hsu, D., & Srinivasa, S. (2020). Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning. *ACM Transactions on Human-Robot Interaction*, *9*(2), 1–23. https://doi.org/10.1145/3359616

Chen, T., Campbell, D., Gonzalez, L. F., & Coppin, G. (2015). Increasing Autonomy Transparency through capability communication in multiple heterogeneous UAV management. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2434–2439. https://ieeexplore.ieee.org/abstract/document/7353707/

Chien, S.-Y., Lewis, M., Semnani-Azad, Z., & Sycara, K. (2014). An Empirical Model of Cultural Factors on Trust in Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1), 859–863. https://doi.org/10.1177/1541931214581181

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between Trust Attitudes Toward Automation, Hofstede's Cultural Dimensions, and Big Five Personality Traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 841–845. https://doi.org/10.1177/1541931213601192

Choi, S. Y. (2018). Agent-Based Human-Robot Interaction Simulation Model for the Analysis of Operator Performance in the Supervisory Control of UGVs. *INTERNATIONAL JOURNAL OF PRECISION ENGINEERING AND MANUFACTURING*, *19*(5), 685–693. https://doi.org/10.1007/s12541-018-0082-3

Chung, A. (2017). *Human Factors and Ergonomics as a Scientific Discipline: The Relationship between Theory, Research, and Practice* [PhD Thesis, UNSW Sydney]. https://unsworks.unsw.edu.au/entities/publication/46538a57-0beb-4b79-bd18-16c97a1b79a2

Cilliers, P. (2008). 3.1 knowing complex systems: The limits of understanding. *A Vision of Transdisciplinarity: Laying Foundations for a World Knowledge Dialogue*, *43*. https://www.google.com/books?hl=no&lr=&id=MbC45UX_4gMC&oi=fnd&pg=PA43&dq=cilliers+1998+complex+systems&ots=CY6R_vqOHr&sig=CVzmu35swuS7rCpLYGxA7zurC5s

Clare, A. S. (2013). *Modeling real-time human-automation collaborative scheduling of unmanned vehicles*. MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF AERONAUTICS AND ASTRONAUTICS.

Cohen, M. S., Parasuraman, R., Serfaty, D., & Andes, R. (1997). Trust in decision aids: A model and a training strategy. *Arlington, VA: Cognitive Technologies, Inc*.

Coito, J. (2021). Maritime autonomous surface ships: New possibilities—and challenges—in ocean law and policy. *International Law Studies*, *97*(1), 19.

Comte, A., & Bridges, J. H. (2015). *A general view of positivism*. Routledge. https://www.taylorfrancis.com/books/mono/10.4324/9781315645780/general-view-positivism-auguste-comte-bridges

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, *58*(6), 737–758.

Coyle, J. M., Exelby, D., & Holt, J. (1999). System dynamics in defence analysis: Some case studies. *Journal of the Operational Research Society*, *50*(4), 372–382. https://doi.org/10.1057/palgrave.jors.2600711

Cramer, K. M. (2013). Six Criteria of a Viable Theory: Putting Reversal Theory to the Test. *Journal of Motivation, Emotion, and Personality: Reversal Theory Studies*, *1*. https://doi.org/10.12689/jmep.2013.102

Crossan, F. (2003). Research philosophy: Towards an understanding. *Nurse Researcher (through 2013)*, *11*(1), 46.

Daly, M. A. (2002). *Task load and automation use in an uncertain environment*. https://scholar.afit.edu/etd/4394/

Davis, J. P., Eisenhardt, K. M., & Bingham, C. B. (2007). Developing Theory Through Simulation Methods. *Academy of Management Review*, *32*(2), 480–499. https://doi.org/10.5465/amr.2007.24351453

De Bot, K. (2017). Complexity theory and dynamic systems theory. *Complexity Theory and Language Development: In Celebration of Diane Larsen-Freeman*, *48*, 51.

De Greene, K. B. (1980). Major conceptualproblems in the systems management of human factors/ergonomics research. *Ergonomics*, *23*(1), 3–11. https://doi.org/10.1080/00140138008924713

De Keyser, V., Decortis, F., & Van Daele, A. (1988). The approach of Francophone ergonomy: Studying new technologies. *The Meaning of Work and Technological Options. London: John Willey & Sons. PMCid: PMC1050468*.

De Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719–735.

de Vries, P. W. (2005). *Trust in systems: Effects of direct and indirect information*. https://research.tue.nl/en/publications/trust-in-systems-effects-of-direct-and-indirect-information

Dekker, S., Cilliers, P., & Hofmeyr, J.-H. (2011). The complexity of failure: Implications of complexity theory for safety investigations. *Safety Science*, *49*(6), 939–945. https://doi.org/10.1016/j.ssci.2011.01.008

Dekker, S., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology & Work*, *6*(2), 79–86. https://doi.org/10.1007/s10111-003-0136-9

Dekker, S. W. (2005). Why we need new accident models. *Contemporary Issues in Human Factors and Aviation Safety*, 181–198.

Dekker, S. W., & Woods, D. D. (2002). MABA-MABA or abracadabra? Progress on human–automation co-ordination. *Cognition, Technology & Work*, *4*(4), 240–244.

Dempsey, P. G., Wogalter, M. S., & Hancock, P. A. (2000). What's in a name? Using terms from definitions to examine the fundamental foundation of human factors and ergonomics science. *Theoretical Issues in Ergonomics Science*, *1*(1), 3–10. https://doi.org/10.1080/146392200308426

Desai, M. (2012). *Modeling trust to improve human-robot interaction* [PhD Thesis, University of Massachusetts Lowell]. https://www.yumpu.com/en/document/read/36708191/modeling-trust-to-improve-human-robot-interaction-umass-lowell-

Descartes, R. (2016). Meditations on first philosophy. In *Seven Masterpieces of Philosophy* (pp. 63–108). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315508818-3/meditations-first-philosophy-ren%C3%A9-descartes

Deutsch, D. (2011). *The Beginning of Infinity: Explanations that Transform The World*. Penguin UK.

Diebold, J. (1952). *Automation: The Advent of the Automatic Factory* (1st edition). Van Nostrand.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. Macmillan International Higher Education.

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

Dore, M. H., & Rosser Jr, J. B. (2007). Do Nonlinear Dynamics in Economics Amount to a Kuhnian Paradigm Shift. *Nonlinear Dynamics, Psychology, and Life Sciences*, *11*(1), 119.

Douglas, L., Aleva, D., & Havig, P. (2007). *Shared displays: An overview of perceptual and cognitive issues*. AIR FORCE RESEARCH LAB WRIGHT-PATTERSON AFB OH WARFIGHTER INTERFACE DIVISION.

Dreyer, L. O., & Oltedal, H. A. (2019). Safety challenges for maritime autonomous surface ships: A systematic review. *The Third Conference on Maritime Human Factors. Haugesund*. https://www.researchgate.net/profile/Leif_Dreyer/publication/372187523_Safety_Challenges_for_Maritime_Autonomous_Surface_Ships_A_Systematic_Review/links/64a824cfb9ed6874a503fa98/Safety-Challenges-for-Maritime-Autonomous-Surface-Ships-A-Systematic-Review.pdf

Drnec, K., Marathe, A. R., Lukos, J. R., & Metcalfe, J. S. (2016). From Trust in Automation to Decision Neuroscience: Applying Cognitive Neuroscience Methods to Understand and Improve Interaction Decisions Involved in Human Automation Interaction. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00290

Drnec, K., Marathe, A. R., Metcalfe, J. S., & Schaefer, K. (2016). The importance of psychophysiological methods in identifying and mitigating degraded situation awareness. *2016 IEEE International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 97–101. https://ieeexplore.ieee.org/abstract/document/7497794/

Dubin, R. (1970). Theory building. *Philosophy and Phenomenological Research*, *31*(2).

Dul, J., Bruder, R., Buckle, P., Carayon, P., Falzon, P., Marras, W. S., Wilson, J. R., & van der Doelen, B. (2012). A strategy for human factors/ergonomics: Developing the discipline and profession. *Ergonomics*, *55*(4), 377–395.

Dybvik, H., Veitch, E., & Steinert, M. (2020). Exploring challenges with designing and developing shore control centers (SCC) for autonomous ships. *Proceedings of the Design Society: DESIGN Conference*, *1*, 847–856. https://www.cambridge.org/core/journals/proceedings-of-the-design-society-design-conference/article/exploring-challenges-with-designing-and-

developing-shore-control-centers-scc-for-autonomous-ships/42B959DDACE34B1790B7CAE2AAE2CA04

Dzindolet, M. T., Beck, H. P., Pierce, L. G., & Dawe, L. A. (2001). A framework of automation use. *Army Research Laboratory. Aberdeen Proving Ground*. https://www.researchgate.net/profile/Hall-Beck/publication/303844546_A_framework_of_automation_use/links/57b1e92c08ae95f9d8f4c36c/A-framework-of-automation-use.pdf

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., Dawe, L. A., & Anderson, B. W. (2001). Predicting Misuse and Disuse of Combat Identification Systems. *Military Psychology*, *13*(3), 147–164. https://doi.org/10.1207/S15327876MP1303_2

Dzitac, I., Filip, F. G., & Manolescu, M.-J. (2017). Fuzzy logic is not fuzzy: World-renowned computer scientist Lotfi A. Zadeh. *International Journal of Computers Communications & Control*, *12*(6), 748–789.

Emad, G. R., & Ghosh, S. (2023). Identifying essential skills and competencies towards building a training framework for future operators of autonomous ships: A qualitative study. *WMU Journal of Maritime Affairs*, *22*(4), 427–445. https://doi.org/10.1007/s13437-023-00310-9

Endsley, M. R. (1996). Automation and situation awareness. *Automation and Human Performance: Theory and Applications*, *20*, 163–181.

Endsley, M. R. (2018). Level of Automation Forms a Key Aspect of Autonomy Design. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 29–34. https://doi.org/10.1177/1555343417723432

Endsley, M. R., & Kaber, D. B. (1999). Level of automation effects on performance, situation awareness and workload in a dynamic control task. *Ergonomics*, *42*(3), 462–492.

Endsley, M. R., & Kiris, E. O. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors*, *37*(2), 381–394.

Evans, A. M., Dillon, K. D., Goldin, G., & Krueger, J. I. (2011). Trust and self-control: The moderating role of the default. *Judgment and Decision Making*, *6*(7), 697–705.

Evans, A. M., & Krueger, J. I. (2011). Elements of trust: Risk and perspective-taking. *Journal of Experimental Social Psychology*, *47*(1), 171–177.

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, *42*(6), 1585–1593.

Fawcett, J. (1988). Conceptual Models and Theory Development. *Journal of Obstetric, Gynecologic & Neonatal Nursing*, *17*(6), 400–403. https://doi.org/10.1111/j.1552-6909.1988.tb00465.x

Fereidunian, A., Lesani, H., Zamani, M. A., Kolarijani, M. A. S., Hassanpour, N., & Mansouri, S. S. (2015). A Complex Adaptive System of Systems Approach to Human–Automation Interaction in Smart Grid. In M. Zhou, H. Li, & M. Weijnen (Eds.), *Contemporary Issues in Systems Science and Engineering* (1st ed., pp. 425–500). Wiley. https://doi.org/10.1002/9781119036821.ch12

Fischhoff, B., & Beyth, R. (1975). I knew it would happen: Remembered probabilities of once—future things. *Organizational Behavior and Human Performance*, *13*(1), 1–16.

Fishbein, M., & Ajzen, I. (1975). *Intention and Behavior: An introduction to theory and research*. Addison-Wesley, Reading, MA.

Fitts, P. M. (1951). *Human engineering for an effective air-navigation and traffic-control system.* https://psycnet.apa.org/record/1952-01751-000

Flach, J. M. (1995). Situation Awareness: Proceed with Caution. *Human Factors*, *37*(1), 149–157. https://doi.org/10.1518/001872095779049480

Flach, J. M. (2012). Complexity: Learning to muddle through. *Cognition, Technology & Work*, *14*(3), 187–197. https://doi.org/10.1007/s10111-011-0201-8

Fleener, M. J., & Merritt, M. L. (2007). Paradigms Lost?. *Nonlinear Dynamics, Psychology, and Life Sciences*, *11*(1), 1.

Flood, R., Jackson, M., Ulrich, W., & Midgley, G. (2003). Systems Thinking-A Studie of Alternatives of. *Int. J. Comput. Syst. Signals*, *4*(1), 33–41.

Fodor, J. A. (1983). *The modularity of mind*. MIT press. https://books.google.com/books?hl=no&lr=&id=0vg0AwAAQBAJ&oi=fnd&pg=PP8&dq=Fodor,+1983&ots=IxDp-ZYXnF&sig=DgwC5igcCsF9-fWrkDe8JbAlj1I

Forrester, J. W. (1987). Lessons from system dynamics modeling. *System Dynamics Review*, *3*(2), 136–149. https://doi.org/10.1002/sdr.4260030205

Forrester, J. W. (1992). Policies, decisions and information sources for modeling. *European Journal of Operational Research*, *59*(1), 42–63.

Forrester, J. W. (1997). Industrial Dynamics. *Journal of the Operational Research Society*, *48*(10), 1037–1041. https://doi.org/10.1057/palgrave.jors.2600946

Forrester, J. W. (2012). Industrial Dynamics: A Major Breakthrough for Decision Makers. In P. Klaus & S. Müller (Eds.), *The Roots of Logistics* (pp. 141–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-27922-5_13

Fosnot, C. T. (2013). *Constructivism: Theory, perspectives, and practice*. Teachers College Press. https://books.google.com/books?hl=no&lr=&id=-pIbAgAAQBAJ&oi=fnd&pg=PT9&dq=constructivism+&ots=tzJ9VTmyxF&sig=My0CIvEseNYeLkkoYTDA9VyL0Xc

Frankfort-Nachmias, C., Nachmias, D., & DeWaard, J. (2014). *Research Methods in the Social Sciences* (Eighth edition). Worth Publishers.

Freedman, D. A. (2010). *Statistical models and causal inference: A dialogue with the social sciences*. Cambridge University Press.

Fuchs, A. (2013). *Nonlinear Dynamics in Complex Systems: Theory and Applications for the Life-, Neuro- and Natural Sciences*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-33552-5

Gao, F., Clare, A. S., Macbeth, J. C., & Cummings, M. L. (2013). Modeling the impact of operator trust on performance in multiple robot control. *AAAI Spring Symposium - Technical Report*, *SS-13-07*, 16–22. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84883414355&partnerID=40&md5=e368acd7ad55074c00ba454ea8eebff2

Gao, J., & Lee, J. D. (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), 943–959.

Gardner, M. P. (1985). Mood states and consumer behavior: A critical review. *Journal of Consumer Research*, *12*(3), 281–300.

Garfinkel, A. (1982). Forms of explanation: Rethinking the questions in social theory. *British Journal for the Philosophy of Science*, *33*(4).

Gauch, H. G. (2012). *Scientific method in brief*. Cambridge University Press.

Gauch Jr, H. G. (2012). *Scientific method in brief*. Cambridge University Press.

Getty, R. L. (1995). Should we view ergonomics as a science, an applied engineering practice or an umbrella multi-discipline program? What is legitimate or illegitimate application of ergonomics? *Advances in Industrial Ergonomics and Safety VII, London: Taylor & Francis*.

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). Extending the Technology Acceptance Model to assess automation. In *COGNITION TECHNOLOGY & WORK* (Vol. 14, Issues 1, SI, pp. 39–49). SPRINGER LONDON LTD. https://doi.org/10.1007/s10111-011-0194-3

Goddard, K., Roudsari, A., & Wyatt, J. C. (2014). Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics*, *83*(5), 368–375.

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96.

Gould, S. (1991). *Ever Since Darwin: Reflections in Natural History*. Penguin Books. https://books.google.com/books/about/Ever_Since_Darwin.html?hl=no&id=hb9k3LXnC6gC

Green, W., & Jordan, P. W. (1999). *Human factors in product design: Current practice and future trends*. CRC Press. https://www.taylorfrancis.com/books/mono/10.1201/9781498702096/human-factors-product-design-patrick-jordan-green

Größler, A., Thun, J.-H., & Milling, P. M. (2008). System Dynamics as a Structural Theory in Operations Management. *Production and Operations Management*, *17*(3), 373–384. https://doi.org/10.3401/poms.1080.0023

Guastello, S. J. (2007). Non-linear dynamics and leadership emergence. *The Leadership Quarterly*, *18*(4), 357–369.

Guastello, S. J. (2017). Nonlinear dynamical systems for theory and research in ergonomics. *Ergonomics*, *60*(2), 167–193. https://doi.org/10.1080/00140139.2016.1162851

Guastello, S. J. (2023). *Human factors engineering and ergonomics: A systems approach*. CRC Press. https://www.taylorfrancis.com/books/mono/10.1201/9781003359128/human-factors-engineering-ergonomics-stephen-guastello

Guastello, S. J., & Liebovitch, L. S. (2009). *Introduction to nonlinear dynamics and complexity.* https://psycnet.apa.org/record/2008-18181-001

Guo, Y., & Yang, X. J. (2020). Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics*. https://doi.org/10.1007/s12369-020-00703-3

Guo, Y., & Yang, X. J. (2021). Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach. *International Journal of Social Robotics*, *13*(8), 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

Haddow, G., & Klobas, J. E. (2004). Communication of research to practice in library and information science: Closing the gap. *Library & Information Science Research*, *26*(1), 29–43.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hancock, P. A., Jagacinski, R. J., Parasuraman, R., Wickens, C. D., Wilson, G. F., & Kaber, D. B. (2013). Human-automation interaction research: Past, present, and future. *Ergonomics in Design*, *21*(2), 9–14.

Hanneman, R. (1988). *Computer-assisted theory building: Modeling dynamic social systems*. Sage Publications.

Hanneman, R. A. (1991). Computer-assisted theory building: Modeling dynamic social systems. *System Dynamics Review*, *7*(2), 202–203. https://doi.org/10.1002/sdr.4260070208

Harré, R. (1984). Personal being: A theory for individual psychology. *Ethics*, *95*(4).

Hartwich, F., Witzlack, C., Beggiato, M., & Krems, J. F. (2019). The first impression counts—A combined driving simulator and test track study on the development of trust and acceptance of highly automated driving. In *TRANSPORTATION RESEARCH PART F-TRAFFIC PSYCHOLOGY AND BEHAVIOUR* (Vol. 65, pp. 522–535). ELSEVIER SCI LTD. https://doi.org/10.1016/j.trf.2018.05.012

Haslam, R., & Waterson, P. (2013). Ergonomics and Sustainability. *Ergonomics*, *56*(3), 343–347. https://doi.org/10.1080/00140139.2013.786555

Heimgärtner, R. (2007). Cultural Differences in Human Computer Interaction: Results from Two Online Surveys. *Isi*, 145–157. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=d245aba653c269ca370657a7a38fb242ce6c66a7#page=157

Hitchcock, C., & Woodward, J. (2003). Explanatory generalizations, part II: Plumbing explanatory depth. *Noûs*, *37*(2), 181–199.

Hockey, G. R. J. (2008). From theory to practice: Commentary on Bartlett (1962). *Ergonomics*, *51*(1), 21–29. https://doi.org/10.1080/00140130701800852

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

Holden, M. T., & Lynch, P. (2004). Choosing the appropriate methodology: Understanding research philosophy. *The Marketing Review*, *4*(4), 397–409.

Hollnagel, E. (2002). Time and time again. *Theoretical Issues in Ergonomics Science*, *3*(2), 143–158. https://doi.org/10.1080/14639220210124111

Hollnagel, E. (2003). *Handbook of Cognitive Task Design*. CRC Press.

Hollnagel, E. (2021). The necessity of muddling through. In *Resilient Health Care* (pp. 9–13). CRC Press. https://www.taylorfrancis.com/chapters/edit/10.1201/9781003095224-3/necessity-muddling-erik-hollnagel

Hollnagel, E., & Woods, D. D. (1983). Cognitive systems engineering: New wine in new bottles. *International Journal of Man-Machine Studies*, *18*(6), 583–600.

Hollnagel, E., & Woods, D. D. (2005). *Joint cognitive systems: Foundations of cognitive systems engineering*. CRC press.

Hoogendoorn, M., Jaffry, S. W., van Maanen, P.-P., & Treur, J. (2013). Modelling biased human trust dynamics. *Web Intelligence and Agent Systems: An International Journal*, *11*(1), 21–40. https://doi.org/10.3233/WIA-130260

Horrobin, D. F. (1969). The Assumptions of Science. In D. F. Horrobin, *Science is God* (pp. 13–19). Springer Netherlands. https://doi.org/10.1007/978-94-011-6106-0_2

Hou, M., Ho, G., & Dunwoody, D. (2021). IMPACTS: A trust model for human-autonomy teaming. *Human-Intelligent Systems Integration*, *3*(2), 79–97.

Hu, W.-L., Akash, K., Reid, T., & Jain, N. (2018). Computational modeling of the dynamics of human trust during human–machine interactions. *IEEE Transactions on Human-Machine Systems*, *49*(6), 485–497.

Hu, W.-L., Akash, K., Reid, T., & Jain, N. (2019). Computational Modeling of the Dynamics of Human Trust During Human-Machine Interactions. In *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS* (Vol. 49, Issue 6, pp. 485–497). IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC. https://doi.org/10.1109/THMS.2018.2874188

Hugh, T. B., & Dekker, S. W. (2009). Hindsight bias and outcome bias in the social construction of medical negligence: A review. *Journal of Law and Medicine*, *16*(5), 846–857.

Hughes, J., & Sharrock, W. (2017). *Theory and methods in sociology: An introduction to sociological thinking and practice*. Bloomsbury Publishing.

https://www.google.com/books?hl=no&lr=&id=JSJHEAAAQBAJ&oi=fnd&pg=PR1&dq=(Hughes+et+al.,+2017+methods&ots=5Uk0PfZjwO&sig=9ji4X6l94kUiIdch1ubzFwqE-rg

Hussein, A., Elsawah, S., & Abbass, H. (2019). A System Dynamics Model for Human Trust in Automation under Speed and Accuracy Requirements. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *63*(1), 822–826. https://doi.org/10.1177/1071181319631167

Hussein, A., Elsawah, S., & Abbass, H. A. (2020). Trust Mediating Reliability-Reliance Relationship in Supervisory Control of Human-Swarm Interactions. In *HUMAN FACTORS* (Vol. 62, Issue 8, pp. 1237–1248). SAGE PUBLICATIONS INC. https://doi.org/10.1177/0018720819879273

Hutchins, E. (1996). *Cognition in the Wild* (Revised ed. edition). A Bradford Book.

Ihde, D. (2002). *Bodies in technology* (Vol. 5). U of Minnesota Press. https://books.google.com/books?hl=no&lr=&id=kLM9gfnPcFAC&oi=fnd&pg=PR9&dq=Ihde,+2002&ots=hmWrZGOUBE&sig=AR_P2VSPgULaZbyAhAJIOw-KpV8

IMO, M. 99/WP. 9. (2018). *Regulatory scoping exercise for the use of Maritime Autonomous Surface Ships (MASS), Report of the working group*.

Israel, G. (2005). The Science of Complexity: Epistemological Problems and Perspectives. *Science in Context*, *18*(3), 479–509. https://doi.org/10.1017/S0269889705000621

Itoh, M. (2011). A model of trust in automation: Why humans over-trust? *SICE Annual Conference 2011*, 198–201.

Jackson, M. C. (2006). Creative holism: A critical systems approach to complex problem situations. *Systems Research and Behavioral Science*, *23*(5), 647–657. https://doi.org/10.1002/sres.799

Jalonen, R., Tuominen, R., & Wahlström, M. (2017). *Safety of Unmanned Ships-Safe Shipping with Autonomous and Remote Controlled Ships*.

Jodlowski, M. T. (2008). *Extending long term working memory theory to dynamic domains: The nature of retrieval structures in situation awareness*. Mississippi State University.

Johannsen, G. (1997). Conceptual design of multi-human machine interfaces. *Control Engineering Practice*, *5*(3), 349–361.

John, O. P., & Srivastava, S. (1999). *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*. http://www.personality-project.org/revelle/syllabi/classreadings/john.pdf

Jokioinen, E., Poikonen, J., Jalonen, R., & Saarni, J. (2016). Remote and autonomous ships-the next steps. *AAWA Position Paper, Rolls Royce Plc, London*.

Jonker, C. M., & Treur, J. (1999). Formal Analysis of Models for the Dynamics of Trust Based on Experiences. In F. J. Garijo & M. Boman (Eds.), *Multi-Agent System Engineering* (Vol. 1647, pp. 221–231). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-48437-X_18

Jordan, S. (2019). Captain, my captain: A look at autonomous ships and how they should operate under admiralty law. *Ind. Int'l & Comp. L. Rev.*, *30*, 283.

Jost, J. (2005). *Dynamical systems: Examples of complex behaviour*. Springer Science & Business Media. https://books.google.com/books?hl=no&lr=&id=kanonbwiqzcC&oi=fnd&pg=PA1&dq=complex+dynamic+systems+theory+nonlinear&ots=LAvxvYYbBe&sig=19YocssOu9UJcyi8Ao5T0QzrWf0

Kahraman, C., Gülbay, M., & Kabak, Ö. (2006). Applications of Fuzzy Sets in Industrial Engineering: A Topical Classification. In C. Kahraman (Ed.), *Fuzzy Applications in Industrial Engineering* (Vol. 201, pp. 1–55). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-33517-X_1

Kampmann, C. E., & Oliva, R. (2008). Structural dominance analysis and theory building in system dynamics. *Systems Research and Behavioral Science*, *25*(4), 505–519. https://doi.org/10.1002/sres.909

Kant, I. (1908). Critique of pure reason. 1781. *Modern Classical Philosophers, Cambridge, MA: Houghton Mifflin*, 370–456.

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioural science*. Chandler Publishing.

Karwowski, W. (2005). Ergonomics and human factors: The paradigms for science, engineering, design, technology and management of human-compatible systems. *Ergonomics*, *48*(5), 436–463. https://doi.org/10.1080/00140130400029167

Karwowski, W. (2012). A Review of Human Factors Challenges of Complex Adaptive Systems: Discovering and Understanding Chaos in Human Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(6), 983–995. https://doi.org/10.1177/0018720812467459

Kazi, T. A., Stanton, N. A., Walker, G. H., & Young, M. S. (2007). Designer driving: Drivers' conceptual models and level of trust in adaptive cruise control. *International Journal of Vehicle Design*, *45*(3), 339. https://doi.org/10.1504/IJVD.2007.014909

Kelly, C., Boardman, M., Goillau, P., & Jeannot, E. (2001). Principles and Guidelines for the Development of Trust in Future ATM Systems: A Literature Review. *European Organisation for the Safety of Air Navigation: 48pp*.

Kenesei, Z., Ásványi, K., Kökény, L., Jászberényi, M., Miskolczi, M., Gyulavári, T., & Syahrivar, J. (2022). Trust and perceived risk: How different manifestations affect the adoption of autonomous vehicles. *Transportation Research Part A: Policy and Practice*, *164*, 379–393. https://doi.org/10.1016/j.tra.2022.08.022

Kennedy, W. G. (2012). Modelling Human Behaviour in Agent-Based Models. In A. J. Heppenstall, A. T. Crooks, L. M. See, & M. Batty (Eds.), *Agent-Based Models of Geographical Systems* (pp. 167–179). Springer Netherlands. https://doi.org/10.1007/978-90-481-8927-4_9

Kerlinger, F. N. (1979). *Behavioral research a conceptual approach*.

Kivunja, C. (2018). Distinguishing between Theory, Theoretical Framework, and Conceptual Framework: A Systematic Review of Lessons from the Field. *International Journal of Higher Education*, *7*(6), 44. https://doi.org/10.5430/ijhe.v7n6p44

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, *12*, 604977. https://doi.org/10.3389/fpsyg.2021.604977

Kongsberg. (2023). *K-Sim—Maritime Simulation—Education, Training and Studies*. https://www.kongsbergdigital.com/resources

Konstandinidou, M., Nivolianitou, Z., Kiranoudis, C., & Markatos, N. (2006). A fuzzy modeling application of CREAM methodology for human reliability analysis. *Reliability Engineering & System Safety*, *91*(6), 706–716.

Körber, M. (2019). Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (Vol. 823, pp. 13–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_2

Kraus, J. M. (2020). *Psychological processes in the formation and calibration of trust in automation* [PhD Thesis]. Universität Ulm.

Kraus, J., Scholz, D., Stiegemeier, D., & Baumann, M. (2020). The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *62*(5), 718–736. https://doi.org/10.1177/0018720819853686

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433.

Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*.

Kroeze, J. H. (2012). Postmodernism, interpretivism, and formal ontologies. In *Research methodologies, innovations and philosophies in software systems engineering and information systems* (pp. 43–62). IGI Global. https://www.igi-global.com/chapter/content/63257

Kuhn, T. S. (1962). Historical Structure of Scientific Discovery: To the historian discovery is seldom a unit event attributable to some particular man, time, and place. *Science*, *136*(3518), 760–764.

Kuhn, T. S. (1977). Second Thoughts on Paradigms. The Essential Tension. *Selected Studies in Scientific Tradition and Change. TS Kuhn. Chicago, Il/London, Chicago University Press*.

Lakatos, I. (1970). *Falsification and the methodology of scientific research programmes. Criticism and the Growth of Knowledge. I. Lakatos and A. Musgrave*. Cambridge, Cambridge University Press.

Lambrou, M., & Ota, M. (2017). Shipping 4.0: Technology stack and digital innovation challenges. *IAME 2017 Conference*, 1–20. https://www.researchgate.net/profile/Maria-Lambrou/publication/320102036_Shipping_40_Technology_Stack_and_Digital_Innovation_Challenges/links/59ce3450458515cc6aaa06fb/Shipping-40-Technology-Stack-and-Digital-Innovation-Challenges.pdf

Lane, D. C. (2001). *Rerum cognoscere causas*: Part I — How do the ideas of system dynamics relate to traditional social theories and the voluntarism/determinism debate? *System Dynamics Review*, *17*(2), 97–118. https://doi.org/10.1002/sdr.209

Lane, D. C. (2015). Validity is a Matter of Confidence-But Not Just in System Dynamics: Validity is a Matter of Confidence. *Systems Research and Behavioral Science*, *32*(4), 450–458. https://doi.org/10.1002/sres.2337

Lane, D. C., & Schwaninger, M. (2008). Theory building with system dynamics: Topic and research contributions. *Systems Research and Behavioral Science*, *25*(4), 439–445. https://doi.org/10.1002/sres.912

Laudan, L. (1978). *Progress and its problems: Towards a theory of scientific growth* (Vol. 282). Univ of California Press.

Laudan, L. (1986). Science and values. In *Science and Values*. University of California Press.

Laurinen, M. (2016). *Remote and Autonomous Ships: The next steps, AAWA: Advanced Autonomous Waterborne Applications*.

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

Lee, J., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Leveson, N. G. (2002). System safety engineering: Back to the future. *Massachusetts Institute of Technology*.

Levine, R. L. (2000). System dynamics applied to psychological and social problems. *Proceedings of the 18th International Conference of the System Dynamics Society*, *126*. https://proceedings.systemdynamics.org/2000/PDFs/levine75.pdf

Levine, R. L., & Doyle, J. K. (2002). Modeling generic structures and patterns in social psychology. *Proceeding of the 20th System Dynamics Conference, Italy*. https://proceedings.systemdynamics.org/2002/proceed/papers/Levin2.pdf

Levine, R. L., & Fitzgerald, H. E. (Eds.). (1992). *Analysis of Dynamic Psychological Systems: Volume 2 Methods and Applications*. Springer US. https://doi.org/10.1007/978-1-4615-6440-9

Lewin, K. (1951). *Field theory in social science: Selected theoretical papers (Edited by Dorwin Cartwright.)*.

Lewis, J. D., & Weigert, A. J. (2012). The social dynamics of trust: Theoretical and empirical research, 1985-2012. *Social Forces*, *91*(1), 25–31.

Li, M., Holthausen, B., Stuck, R., & Walker, B. (2019). *No Risk No Trust: Investigating Perceived Risk in Highly Automated Driving* (p. 185). https://doi.org/10.1145/3342197.3344525

Liang, F., Brunelli, M., & Rezaei, J. (2020). Consistency issues in the best worst method: Measurements and thresholds. *Omega*, *96*, 102175.

Licklider, J. C. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics*, *1*, 4–11.

Lipton, P. (1990). Contrastive Explanation. *Royal Institute of Philosophy Supplement*, *27*, 247–266. https://doi.org/10.1017/S1358246100005130

Locke, J. (1847). *An essay concerning human understanding*. Kay & Troutman. https://www.google.com/books?hl=no&lr=&id=2aw8AAAAYAAJ&oi=fnd&pg=PA1&dq=John+Locke+An+Essay+Concerning+Human+Understanding&ots=Zb62n477SG&sig=waxGKEOfJB6dMl7P61ciXb4QL9o

Lu, Y., & Sarter, N. (2019). Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability. *IEEE Transactions on Human-Machine Systems*, *49*(6), 560–568. https://doi.org/10.1109/THMS.2019.2930980

Luna-Reyes, L. F., & Andersen, D. L. (2003). Collecting and analyzing qualitative data for system dynamics: Methods and models: Collecting and Analyzing Qualitative Data. *System Dynamics Review*, *19*(4), 271–296. https://doi.org/10.1002/sdr.280

Lund, I. O., & Rundmo, T. (2009). Cross-cultural comparisons of traffic safety, risk perception, attitudes and behaviour. *Safety Science*, *47*(4), 547–553.

Ma, R. (2005). *The effect of in-vehicle automation and reliability on driver situation awareness and trust*. North Carolina State University. https://search.proquest.com/openview/2abbbed4b7ac6186e43305412b842693/1?pq-origsite=gscholar&cbl=18750&diss=y

Madhavan, P., & Wiegmann, D. A. (2004). A new look at the dynamics of human-automation trust: Is trust in humans comparable to trust in machines? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *48*(3), 581–585.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human Factors*, *48*(2), 241–256.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th Australasian Conference on Information Systems*, *53*, 6–8.

Mallam, S. C., Nazir, S., & Sharma, A. (2020). The human element in future Maritime Operations – perceived impact of autonomous shipping. *Ergonomics*, *63*(3), 334–345. https://doi.org/10.1080/00140139.2019.1659995

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

Marchionni, C. (2012). Geographical Economics and its Neighbours—Forces Towards and Against Unification. In *Philosophy of Economics* (pp. 425–458). Elsevier. https://doi.org/10.1016/B978-0-444-51676-3.50015-4

Markus, M. L., & Robey, D. (1988). Information Technology and Organizational Change: Causal Structure in Theory and Research. *Management Science*, *34*(5), 583–598. https://doi.org/10.1287/mnsc.34.5.583

Masalonis, A. J., Duley, J. A., Galster, S. M., Castano, D. J., Metzger, U., & Parasuraman, R. (1998). Air Traffic Controller Trust in a Conflict Probe during Free Flight. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *42*(23), 1601–1601. https://doi.org/10.1177/154193129804202304

Masalonis, A. J., & Parasuraman, R. (1999). Trust as a construct for evaluation of automated aids: Past and future theory and research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*(3), 184–187.

Matthews, G., Lin, J., Panganiban, A. R., & Long, M. D. (2020). Individual Differences in Trust in Autonomous Robots: Implications for Transparency. In *IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS* (Vol. 50, Issues 3, SI, pp. 234–244). IEEE-INST ELECTRICAL ELECTRONICS ENGINEERS INC. https://doi.org/10.1109/THMS.2019.2947592

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model Of Organizational Trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335

McBride, S. E., Rogers, W. A., & Fisk, A. D. (2011). Understanding the Effect of Workload on Automation Use for Younger and Older Adults. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(6), 672–686. https://doi.org/10.1177/0018720811421909

McCrae, R. R., & John, O. P. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

McElreath, R. (2018). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Chapman and Hall/CRC.

Meadows, D. H. (1989). System dynamics meets the press. *System Dynamics Review*, *5*(1), 69–80. https://doi.org/10.1002/sdr.4260050106

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115. https://doi.org/10.1086/288135

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*(4), 806.

Meister, D. (2000). Theoretical issues in general and developmental ergonomics. *Theoretical Issues in Ergonomics Science*, *1*(1), 13–21. https://doi.org/10.1080/146392200308444

Meister, D. (2018). *The history of human factors and ergonomics*. CRC Press.

Meister, D., & Enderwick, T. P. (2001). *Human factors in system design, development, and testing*. CRC Press. https://books.google.com/books?hl=no&lr=&id=VNlhXpMEu1gC&oi=fnd&pg=PP1&dq=meister+human+factors&ots=3BlidTTvlv&sig=lsqrTJKdNh-adIIj-Z5qZFp_0c0

Meleis, A. I. (2012). A model for evaluation of theories: Description, analysis, critique, testing, and support. *Theoretical Nursing: Development and Progress*, 179–206.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors*, *53*(4), 356–370.

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I Trust It, but I Don't Know Why: Effects of Implicit Attitudes Toward Automation on Trust in an Automated System. In *HUMAN FACTORS* (Vol. 55, Issue 3, pp. 520–534). SAGE PUBLICATIONS INC. https://doi.org/10.1177/0018720812465081

Merritt, S. M., Huber, K., LaChapell-Unnerstall, J., & Lee, D. (2014). Continuous calibration of trust in automated systems. *Air Force Research Laboratory Technical Report AFRL-RH-WP-TR-2014-0026*. https://apps.dtic.mil/sti/citations/tr/ADA606748

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210. https://doi.org/10.1518/001872008X288574

Meshkati, N. (1991). Human factors in large-scale technological systems' accidents: Three Mile Island, Bhopal, Chernobyl. *Industrial Crisis Quarterly*, *5*(2), 133–154. https://doi.org/10.1177/108602669100500203

Metzger, U., & Parasuraman, R. (2005). Automation in Future Air Traffic Management: Effects of Decision Aid Reliability on Controller Performance and Mental Workload. *Human Factors*, 15.

Miller, C. A. (2021). Trust, transparency, explanation, and planning: Why we need a lifecycle perspective on human-automation interaction. In *Trust in human-robot interaction* (pp. 233–257). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780128194720000113

Miller, C., Funk, H., Wu, P., Goldman, R., Meisner, J., & Chapman, M. (2005). The Playbook™ approach to adaptive automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *49*, 15–19.

Mitchell, M. (2006). Complex systems: Network thinking. *Artificial Intelligence*, *170*(18), 1194–1212.

Mitroff, I. I. (1969). Fundamental Issues in the Simulation of Human Behavior: A Case in the Strategy of Behavioral Science. *Management Science*, *15*(12), B-635-B-649. https://doi.org/10.1287/mnsc.15.12.B635

Moray, N. (2000). Culture, politics and ergonomics. *Ergonomics*, *43*(7), 858–868. https://doi.org/10.1080/001401300409062

Moray, N. (2008). The Good, the Bad, and the Future: On the Archaeology of Ergonomics. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 411–417. https://doi.org/10.1518/001872008X288439

Moray, N., Boff, K. R., Kaufmann, L., & Thomas, J. P. (1986). *Handbook of Perception and Human Performance Cognitive Processes and Performance: Vol. II* (null, Ed.).

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control*, *21*(4–5), 203–211.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. In *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-APPLIED* (Vol. 6, Issue 1, pp. 44–58). AMER PSYCHOLOGICAL ASSOC. https://doi.org/10.1037//0278-7393.6.1.44

Morton, A. (1990). Mathematical modelling and contrastive explanation. *Canadian Journal of Philosophy Supplementary Volume*, *16*, 251–270.

Mosier, K. L., & Skitka, L. J. (1999). Automation Use and Automation Bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*(3), 344–348. https://doi.org/10.1177/154193129904300346

Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1998). Automation bias: Decision making and performance in high-tech cockpits. *The International Journal of Aviation Psychology*, *8*(1), 47–63.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5–6), 527–539.

Muir, B. M. (1994a). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922.

Muir, B. M. (1994b). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, *37*(11), 1905–1922.

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. https://doi.org/10.1080/00140139608964474

Nahavandi, S. (2017). Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems. *IEEE Systems, Man, and Cybernetics Magazine*, *3*(1), 10–17.

Nam, C., Walker, P., Li, H., Lewis, M., & Sycara, K. (2020). Models of Trust in Human Control of Swarms With Varied Levels of Autonomy. *IEEE Transactions on Human-Machine Systems*, *50*(3), 194–204. https://doi.org/10.1109/THMS.2019.2896845

Nass, C., Fogg, B. J., & Moon, Y. (1996). Can computers be teammates? *International Journal of Human-Computer Studies*, *45*(6), 669–678. https://doi.org/10.1006/ijhc.1996.0073

Naylor, T. H., & Finger, J. M. (1967). Verification of Computer Simulation Models. *Management Science*, *14*(2), B-92-B-101. https://doi.org/10.1287/mnsc.14.2.B92

Ng, H. K., Grabbe, S., & Mukherjee, A. (2009, August 10). Design and Evaluation of a Dynamic Programming Flight Routing Algorithm Using the Convective Weather Avoidance Model. *AIAA Guidance, Navigation, and Control Conference*. AIAA Guidance, Navigation, and Control Conference, Chicago, Illinois. https://doi.org/10.2514/6.2009-5862

Nickerson, J. V., & Reilly, R. R. (2004). A model for investigating the effects of machine autonomy on human behavior. *37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of The*, 10 pp. https://doi.org/10.1109/HICSS.2004.1265325

Niederée, U., Jipp, M., Teegen, U., & Vollrath, M. (2012a). Effects of Observability, Mood States, and Workload on Human Handling Errors When Monitoring Aircraft Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*(1), 1481–1485. https://doi.org/10.1177/1071181312561414

Niederée, U., Jipp, M., Teegen, U., & Vollrath, M. (2012b). *Effects of Warning Strategy on Reaction Times in Different Workload Conditions*. https://elib.dlr.de/76188/

Niiniluoto, I. (1999). *Critical scientific realism*.

Niiniluoto, I. (2017). Optimistic realism about scientific progress. *Synthese*, *194*(9), 3291–3309.

NORMAN, D. A. (1989). *The problem of automation: Inappropriate feedback and interaction not overautomation: Vol. null* (null, Ed.).

Norris, J. N. (2018). Human Factors in Military Maritime and Expeditionary Settings: Opportunity for Autonomous Systems? In J. Chen (Ed.), *Advances in Human Factors in Robots and Unmanned Systems* (Vol. 595, pp. 139–147). Springer International Publishing. https://doi.org/10.1007/978-3-319-60384-1_14

Nudds, T. D., & Villard, M.-A. (2006). Basic Science, Applied Science, and the Radical Middle Ground Science fondamentale, science appliquée et le radicalisme de l'entre-deux. *Avian Conservation and Ecology-Écologie et Conservation Des Oiseaux*, *1*(1), 1.

Olaya, C. (2009). *System Dynamics Philosophical Background and Underpinnings.* https://www.academia.edu/download/40333990/System_Dynamics_Philosophical_Background_and_Underpinnings_-_Olaya2009.pdf

Orasanu, J., & Martin, L. (1998). Errors in aviation decision making: A factor in accidents and incidents. *Proceedings of the Workshop on Human Error, Safety, and Systems Development*, 100–107. https://www.dcs.gla.ac.uk/~johnson/papers/seattle_hessd/judithlynne-p.pdf

Ososky, S., Sanders, T., Jentsch, F., Hancock, P., & Chen, J. Y. (2014). Determinants of system transparency and its influence on trust in and reliance on unmanned robotic systems. *Unmanned Systems Technology XVI*, *9084*, 112–123. https://www.spiedigitallibrary.org/conference-proceedings-of-spie/9084/90840E/Determinants-of-system-transparency-and-its-influence-on-trust-in/10.1117/12.2050622.short

Ottino, J. M. (2003). Complex systems. *American Institute of Chemical Engineers. AIChE Journal*, *49*(2), 292.

Øvergård, K. I. (2008). *Human Co-Agency withTechnical Systems:Investigations of ModellingFrameworks* [Doctoral thesis, Norges teknisk-naturvitenskapelige universitet, Fakultet for samfunnsvitenskap og teknologiledelse, Psykologisk institutt]. https://ntnuopen.ntnu.no/ntnu-xmlui/handle/11250/270343

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, *55*(9), 1059–1072. https://doi.org/10.1080/00140139.2012.691554

Pan, W., Xie, X., He, P., Bao, T., & Li, M. (2021). An automatic route design algorithm for intelligent ships based on a novel environment modeling method. *Ocean Engineering*, *237*, 109603.

Pan, X., Han, C. S., Dauber, K., & Law, K. H. (2007). A multi-agent based framework for the simulation of human and social behaviors during emergency evacuations. *AI & SOCIETY*, *22*(2), 113–132. https://doi.org/10.1007/s00146-007-0126-1

Parasuraman, R. (2000). Designing automation for human use: Empirical studies and quantitative models. *Ergonomics*, *43*(7), 931–951. https://doi.org/10.1080/001401300409125

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410.

Parasuraman, R., & Miller, C. A. (2004). Trust and etiquette in high-criticality automated systems. *Communications of the ACM*, *47*(4), 51. https://doi.org/10.1145/975817.975844

Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced'complacency'. *The International Journal of Aviation Psychology*, *3*(1), 1–23.

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

Parasuraman, R., & Rizzo, M. (2008). *Neuroergonomics: The brain at work* (Vol. 3). Oxford University Press.

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *30*(3), 286–297.

Payne, G. T., Pearson, A. W., & Carr, J. C. (2017). Process and Variance Modeling: Linking Research Questions to Methods in Family Business Research. *Family Business Review*, *30*(1), 11–18. https://doi.org/10.1177/0894486516679749

Perrow, C. (1999). *Normal accidents: Living with high risk technologies*. Princeton university press.

Peters, J. R., Srivastava, V., Taylor, G. S., Surana, A., Eckstein, M. P., & Bullo, F. (2015). Human supervisory control of robotic teams: Integrating cognitive modeling with engineering design. *IEEE Control Systems Magazine*, *35*(6), 57–80.

Petit, J., Dugué, B., & Daniellou, F. (2011). Ergonomic Interventions on Psychosocial Risks in Firms: Theoretical and Methodological Opportunities. *Le Travail Humain*, *74*(4), 391–409.

Poole, M. S., Ven, A. H. V. de, Dooley, K., & Holmes, M. E. (2000). *Organizational Change and Innovation Processes: Theory and Methods for Research*. Oxford University Press.

Poornikoo, M., & Mansouri, M. (2023). Systems approach to modeling controversy in Human factors and ergonomics (HFE). *2023 18th Annual System of Systems Engineering Conference (SoSe)*, 1–8. https://doi.org/10.1109/SoSE59841.2023.10178634

Poornikoo, M., & Øvergård, K. I. (2022). Levels of automation in maritime autonomous surface ships (MASS): A fuzzy logic approach. *Maritime Economics & Logistics*. https://doi.org/10.1057/s41278-022-00215-z

Poornikoo, M., & Øvergård, K. I. (2023). Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation. *Theoretical Issues in Ergonomics Science*, *0*(0), 1–37. https://doi.org/10.1080/1463922X.2023.2233591

Popper, K. (1969). *Conjectures and refutations: The growth of scientific knowledge*. routledge.

Popper, K. (1972). *The logic of scientific discovery*. Hutchinson.

Porathe, T., Fjortoft, K., & Bratbergsengen, I.-L. (2020). Human Factors, autonomous ships and constrained coastal navigation. *IOP Conference Series: Materials Science and Engineering*, *929*(1), 012007. https://iopscience.iop.org/article/10.1088/1757-899X/929/1/012007/meta

Porathe, T., Hoem, Å., Rødseth, Ø., Fjørtoft, K., & Johnsen, S. O. (2018). At least as safe as manned shipping? Autonomous shipping, safety and "human error." In *Safety and Reliability–Safe Societies in a Changing World* (pp. 417–425). CRC Press. https://www.taylorfrancis.com/chapters/oa-edit/10.1201/9781351174664-52/least-safe-manned-shipping-autonomous-shipping-safety-human-error-porathe-hoem-r%C3%B8dseth-fj%C3%B8rtoft-johnsen

Porathe, T., Prison, J., & Man, Y. (2014). Situation awareness in remote control centres for unmanned ships. *Proceedings of Human Factors in Ship Design & Operation, 26-27 February 2014, London, UK*, 93.

Pruyt, E. (2006). What is system dynamics? A paradigmatic inquiry. *Proceedings of the 2006 Conference of the System Dynamics Society*, *29*. https://proceedings.systemdynamics.org/2006/proceed/papers/PRUYT177.pdf

Rajaonah, B., Anceaux, F., & Vienne, F. (2006). Trust and the use of adaptive cruise control: A study of a cut-in situation. *Cognition, Technology & Work*, *8*(2), 146–155. https://doi.org/10.1007/s10111-006-0030-3

Relling, T., Lützhöft, M., Ostnes, R., & Hildre, H. P. (2018). A Human Perspective on Maritime Autonomy. In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Augmented Cognition: Users and Contexts* (Vol. 10916, pp. 350–362). Springer International Publishing. https://doi.org/10.1007/978-3-319-91467-1_27

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95.

Rezaei, J. (2015). Best-worst multi-criteria decision-making method. *Omega*, *53*, 49–57.

Richardson, G. P. (2011). Reflections on the foundations of system dynamics. *System Dynamics Review*, *27*(3), 219–243. https://doi.org/10.1002/sdr.462

Riley, V. (1989). A General Model of Mixed-Initiative Human-Machine Systems. *Proceedings of the Human Factors Society Annual Meeting*, *33*(2), 124–128. https://doi.org/10.1177/154193128903300227

Riley, V. (1996). Operator reliance on automation: Theory and data. *Automation and Human Performance: Theory and Applications*, 19–35.

Rimnac, C. M., & Leopold, S. S. (2014). Basic science, applied science, and product testing. In *Clinical Orthopaedics and Related Research®* (Vol. 472, Issue 8, pp. 2311–2312). LWW. https://journals.lww.com/clinorthop/fulltext/2014/08000/Editorial__Basic_Science,_Applied_Science,_and.1.aspx

Ringbom, H., Viljanen, M., Poikonen, J., & Ilvessalo, S. (n.d.). *Charting Regulatory Frameworks for Maritime Autonomous Surface Ship Testing, Pilots, and Commercial Deployments*. 233.

Roberts, E. B. (1978). Managerial applications of system dynamics. *(No Title)*. https://cir.nii.ac.jp/crid/1130000794816593280

Robinson, S. (2005). Discrete-event simulation: From the pioneers to the present, what next? *Journal of the Operational Research Society*, *56*(6), 619–629. https://doi.org/10.1057/palgrave.jors.2601864

Rodseth, O. J. (2017). From concept to reality: Unmanned merchant ship research in Norway. *2017 IEEE Underwater Technology (UT)*, 1–10. https://doi.org/10.1109/UT.2017.7890328

Rødseth, Ø. J., Lien Wennersberg, L. A., & Nordahl, H. (2021). Towards approval of autonomous ship systems by their operational envelope. *Journal of Marine Science and Technology*. https://doi.org/10.1007/s00773-021-00815-z

Rodseth, O., Nordahl, H., & Hoem, A. (2018). Characterization of Autonomy in Merchant Ships. *2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO)*, 1–7. https://doi.org/10.1109/OCEANSKOBE.2018.8559061

Rorty, R. (1979). Transcendental Arguments, Self-Reference, and Pragmatism. In P. Bieri, R.-P. Horstmann, & L. Krüger (Eds.), *Transcendental Arguments and Science* (pp. 77–103). Springer Netherlands. https://doi.org/10.1007/978-94-009-9410-2_7

Ross, T. J. (2004). *Fuzzy logic with engineering applications* (Vol. 2). Wiley Online Library.

Ross, T. J. (2005). *Fuzzy Logic with Engineering Applications*. John Wiley & Sons.

Rovira, E., McGarry, K., & Parasuraman, R. (2007). Effects of Imperfect Automation on Decision Making in a Simulated Command and Control Task. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *49*(1), 76–87. https://doi.org/10.1518/001872007779598082

Rovira, E., & Parasuraman, R. (2010). Transitioning to future air traffic management: Effects of imperfect automation on controller attention and performance. *Human Factors*, *52*(3), 411–425.

Sadrfaridpour, B., Saeidi, H., Burke, J., Madathil, K., & Wang, Y. (2016). Modeling and Control of Trust in Human-Robot Collaborative Manufacturing. In R. Mittu, D. Sofge, A. Wagner, & W. F. Lawless (Eds.), *Robust Intelligence and Trust in Autonomous Systems* (pp. 115–141). Springer US. https://doi.org/10.1007/978-1-4899-7668-0_7

Salas, E. (2008). At the turn of the 21st century: Reflections on our science. *Human Factors*, *50*(3), 351–353.

Sanchez, J., Rogers, W. A., Fisk, A. D., & Rovira, E. (2014). Understanding reliance on automation: Effects of error type, error distribution, age and experience. *Theoretical Issues in Ergonomics Science*, *15*(2), 134–160. https://doi.org/10.1080/1463922X.2011.611269

Sanders, M. S., & McCormick, E. J. (1998). Human factors in engineering and design. *Industrial Robot: An International Journal*.

Sarter, N. B., & Woods, D. D. (1997). Team Play with a Powerful and Independent Agent: Operational Experiences and Automation Surprises on the Airbus A-320. *Human Factors:*

The Journal of the Human Factors and Ergonomics Society, *39*(4), 553–569. https://doi.org/10.1518/001872097778667997

Saunders, M. N., Lewis, P., Thornhill, A., & Bristow, A. (2015). *Understanding research philosophy and approaches to theory development*. https://oro.open.ac.uk/53393/

Schaefer, K. E., Billings, D. R., Szalma, J. L., Adams, J. K., Sanders, T. L., Chen, J. Y., & Hancock, P. A. (2014). *A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction*. Army Research Lab Aberdeen Proving Ground Md Human Research And Engineering ….

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schwaninger, M., & Grösser, S. (2008). System dynamics as model-based theory building. *Systems Research and Behavioral Science*, *25*(4), 447–465. https://doi.org/10.1002/sres.914

Senge, P. M., & Forrester, J. W. (1980). Tests for building confidence in system dynamics models. *System Dynamics, TIMS Studies in Management Sciences*, *14*, 209–228.

Seong, Y., & Bisantz, A. M. (2000). Modeling human trust in complex, automated systems using a lens model approach. *Automation Technology and Human Performance: Current Research and Trends*, 95–100.

Sepehri, A., Vandchali, H. R., Siddiqui, A. W., & Montewka, J. (2022). The impact of shipping 4.0 on controlling shipping accidents: A systematic literature review. *Ocean Engineering*, *243*, 110162.

Seppelt, B. D., & Lee, J. D. (2019). Keeping the driver in the loop: Dynamic feedback to support appropriate use of imperfect vehicle control automation. *International Journal of Human-Computer Studies*, *125*, 66–80. https://doi.org/10.1016/j.ijhcs.2018.12.009

Shen, S., & Neyens, D. M. (2014). Assessing drivers' performance when automated driver support systems fail with different levels of automation. *Proceedings of the Human Factors and*

Ergonomics Society Annual Meeting, *58*(1), 2068–2072. https://doi.org/10.1177/1541931214581435

Sheridan, T. B. (2017). *Modeling human-system interaction: Philosophical and methodological considerations, with examples*. John Wiley & Sons.

Sheridan, T. B. (2019). Extending Three Existing Models to Analysis of Trust in Automation: Signal Detection, Statistical Parameter Estimation, and Model-Based Control. *Human Factors*, *61*(7), 1162–1170. https://doi.org/10.1177/0018720819829951

Sheridan, T. B. (2021). HUMAN SUPERVISORY CONTROL OF AUTOMATION. In G. Salvendy & W. Karwowski (Eds.), *HANDBOOK OF HUMAN FACTORS AND ERGONOMICS* (1st ed., pp. 736–760). Wiley. https://doi.org/10.1002/9781119636113.ch28

Sheridan, T. B., & Hennessy, R. T. (1984). Research and modeling of supervisory control behavior. Report of a workshop. *Washington DC*. https://apps.dtic.mil/sti/citations/tr/ADA149621

Sheridan, T. B., Sheridan, T. B., Maschinenbauingenieur, K., Sheridan, T. B., & Sheridan, T. B. (2002). *Humans and automation: System design and research issues* (Vol. 280). Human Factors and Ergonomics Society Santa Monica, CA. https://www.cambridge.org/core/services/aop-cambridge-core/content/view/S0263574702274858

Sheridan, T. B., & Verplank, W. L. (1978). *Human and Computer Control of Undersea Teleoperators*. MASSACHUSETTS INST OF TECH CAMBRIDGE MAN-MACHINE SYSTEMS LAB. https://apps.dtic.mil/docs/citations/ADA057655

Sheridan, T. B., Verplank, W. L., & Brooks, T. L. (1978). Human/computer control of undersea teleoperators. *NASA. Ames Res. Center The 14th Ann. Conf. on Manual Control*. https://ntrs.nasa.gov/citations/19790007441

Silva, M. C. (1986). Research testing nursing theory: State of the art. *Advances in Nursing Science*.

Simpson, A., Brander, G. N., & Portsdown, D. R. A. (1995). Seaworthy trust: Confidence in automated data fusion. *The Human-Electronic Crew: Can We Trust the Team*, 77–81.

Sind-Prunier, P. (1996). Bridging the research/practice gap: Human factors practitioners' opportunity for input to define research for the rest of the decade. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *40*(17), 865–867.

Singleton, W. T. (1994). A Personal View: Are researchers producing work that's usable in system design? Maybe not. *Ergonomics in Design: The Quarterly of Human Factors Applications*, *2*(3), 30–34. https://doi.org/10.1177/106480469400200307

Skitka, L. J., Mosier, K., & Burdick, M. D. (2000). Accountability and automation bias. *International Journal of Human-Computer Studies*, *52*(4), 701–717.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human-Computer Studies*, *51*(5), 991–1006.

Smith, S., Booth, K., & Zalewski, M. (1996). *International theory: Positivism and beyond*. Cambridge University Press. https://books.google.com/books?hl=no&lr=&id=pfvD4KaO5vUC&oi=fnd&pg=PR9&dq=Positivism&ots=8uRI6G3RcY&sig=-KSPSa39yVeSquuUdc-HVWdKYDE

Snook, S. H., & Irvine, C. H. (1969). Psychophysical Studies of Physiological Fatigue Criteria. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *11*(3), 291–299. https://doi.org/10.1177/001872086901100311

Society of Automotive Engineers, S. (2014). *Taxonomy and definitions for terms related to on-road motor vehicle automated driving systems (Standard J3016_201401)*. Warrendale, SAE International. https://www.google.com/search?client=firefox-b-d&q=Society+of+Automotive+Engineers.+%282014%29.+Taxonomy+and+defini-+tions+for+terms+related+to+on-road+motor+vehicle+automated+driving+systems+%28Standard+J3016_201401%29.+Warrendale%2C+PA%3A+SAE+International.

Sprott, J. C. (2003). *Chaos and time-series analysis*. Oxford university press. https://academic.oup.com/book/52822

Sterman, J. D. (1994). Learning in and about complex systems. *System Dynamics Review*, *10*(2–3), 291–330. https://doi.org/10.1002/sdr.4260100214

Sterman, J. D. (2000). Business Dynamics, S. *Massachusetts: Jeffrey J. Shelstad*, *196*, 199–201.

Størkersen, K. V. (2021). Safety management in remotely controlled vessel operations. *Marine Policy*, *130*, 104349.

Stramler, J. (1992). Occupational ergonomics in space. *NASA. Johnson Space Center, The Second Conference on Lunar Bases and Space Activities of the 21st Century, Volume 2*. https://ntrs.nasa.gov/citations/19930004825

Sullivan, B. P., Desai, S., Sole, J., Rossi, M., Ramundo, L., & Terzi, S. (2020). Maritime 4.0–opportunities in digitalization and advanced manufacturing for vessel development. *Procedia Manufacturing*, *42*, 246–253.

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied*, *17*(2), 71–96. https://doi.org/10.1037/a0024170

Taylor, F. W. (1911). *The Principles of Scientific Management*. McMaster University Archive for the History of Economic Thought.

Thanh, N. D., Ali, M., & Son, L. H. (2017). A Novel Clustering Algorithm in a Neutrosophic Recommender System for Medical Diagnosis. *Cognitive Computation*, *9*(4), 526–544. https://doi.org/10.1007/s12559-017-9462-8

Thatcher, A., Nayak, R., & Waterson, P. (2020). Human factors and ergonomics systems-based tools for understanding and addressing global problems of the twenty-first century. *Ergonomics*, *63*(3), 367–387. https://doi.org/10.1080/00140139.2019.1646925

Thelen, E. (2005). Dynamic Systems Theory and the Complexity of Change. *Psychoanalytic Dialogues*, *15*(2), 255–283. https://doi.org/10.1080/10481881509348831

Thieme, C. A., Utne, I. B., & Haugen, S. (2018). Assessing ship risk model applicability to Marine Autonomous Surface Ships. *Ocean Engineering*, *165*, 140–154. https://doi.org/10.1016/j.oceaneng.2018.07.040

Thombre, S., Zhao, Z., Ramm-Schmidt, H., García, J. M. V., Malkamäki, T., Nikolskiy, S., Hammarberg, T., Nuortie, H., Bhuiyan, M. Z. H., & Särkkä, S. (2020). Sensors and AI

techniques for situational awareness in autonomous ships: A review. *IEEE Transactions on Intelligent Transportation Systems*, *23*(1), 64–83.

Toulmin, S. (1977). From form to function: Philosophy and history of science in the 1950s and now. *Daedalus*, 143–162.

Turner, S. (2002). *Brains/practices/relativism: Social theory after cognitive science*. University of chicago Press. https://www.google.com/books?hl=no&lr=&id=4sYG0Fj9Y1MC&oi=fnd&pg=PP13&dq=Turner+(2002)+science&ots=dy1DRWKdri&sig=uwJuP5c39q2hfrR0aj40aKfPk14

Urry, J. (2006). Complexity. *Theory, Culture & Society*, *23*(2–3), 111–115. https://doi.org/10.1177/0263276406062818

Utne, I. B., Sørensen, A. J., & Schjølberg, I. (2017). Risk management of autonomous marine systems and operations. *ASME 2017 36th International Conference on Ocean, Offshore and Arctic Engineering*, V03BT02A020-V03BT02A020.

Vagia, M., Transeth, A. A., & Fjerdingen, S. A. (2016). A literature review on the levels of automation during the years. What are the different taxonomies that have been proposed? *Applied Ergonomics*, *53*, 190–202. https://doi.org/10.1016/j.apergo.2015.09.013

Van de Ven, A. H. (2007). *Engaged scholarship: A guide for organizational and social research*. Oxford University Press on Demand.

Van Fraassen, B. C. (1980). *The scientific image*. Oxford University Press.

Van Lange, P. A. M. (2013). What We Should Expect From Theories in Social Psychology: Truth, Abstraction, Progress, and Applicability As Standards (TAPAS). *Personality and Social Psychology Review*, *17*(1), 40–55. https://doi.org/10.1177/1088868312453088

Van Leekwijck, W., & Kerre, E. E. (1999). Defuzzification: Criteria and classification. *Fuzzy Sets and Systems*, *108*(2), 159–178.

Velicer, W. F., Cumming, G., Fava, J. L., Rossi, J. S., Prochaska, J. O., & Johnson, J. (2008). Theory testing using quantitative predictions of effect size. *Applied Psychology*, *57*(4), 589–608.

Verberne, F. M. F., Ham, J., & Midden, C. J. H. (2012). Trust in Smart Systems: Sharing Driving Goals and Giving Information to Increase Trustworthiness and Acceptability of Smart

Systems in Cars. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *54*(5), 799–810. https://doi.org/10.1177/0018720812443825

Veritas, B. (2019). *Guidelines for autonomous shipping. Guidance Note NI 641 DT R01 E.* http://erules.veristar.com/dy/data/bv/pdf/641-NI_2019-10.pdf

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work*. CRC Press.

Wacker, J. G. (2004). A theory of formal conceptual definitions: Developing theory-building measurement instruments. *Journal of Operations Management*, *22*(6), 629–650.

Walker, G. H., Salmon, P. M., Bedinger, M., & Stanton, N. A. (2017). Quantum ergonomics: Shifting the paradigm of the systems agenda. *Ergonomics*, *60*(2), 157–166. https://doi.org/10.1080/00140139.2016.1231840

Walliser, J. C. (2011). *Trust in automated systems the effect of automation level on trust calibration* [PhD Thesis, Monterey, California. Naval Postgraduate School]. https://www.researchgate.net/profile/James-Walliser/publication/235181550_Trust_in_Automated_Systems_The_Effect_of_Automation_Level_on_Trust_Calibration/links/6330564f6063772afd8fe088/Trust-in-Automated-Systems-The-Effect-of-Automation-Level-on-Trust-Calibration.pdf

Wang, L., Jamieson, G. A., & Hollands, J. G. (2008). Selecting Methods for the Analysis of Reliance on Automation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *52*(4), 287–291. https://doi.org/10.1177/154193120805200419

Wang, L., Jamieson, G. A., & Hollands, J. G. (2009). Trust and Reliance on an Automated Combat Identification System. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *51*(3), 281–291. https://doi.org/10.1177/0018720809338842

Waterson, P. (2011). World War II and other historical influences on the formation of the Ergonomics Research Society. *Ergonomics*, *54*(12), 1111–1129. https://doi.org/10.1080/00140139.2011.622796

Waterson, P., & Eason, K. (2009). '1966 and all that': Trends and developments in UK ergonomics during the 1960s. *Ergonomics*, *52*(11), 1323–1341. https://doi.org/10.1080/00140130903229561

Wellman, M. P., Breese, J. S., & Goldman, R. P. (1992). From knowledge bases to decision models. *The Knowledge Engineering Review*, *7*(1), 35–53.

Wertheimer, M. (2012). *Investigations on Gestalt principles*. MIT Press: London, UK. https://books.google.com/books?hl=no&lr=&id=Xd_xCwAAQBAJ&oi=fnd&pg=PA127&dq=In+accordance+with+Gestalt+principles,+the+system+as+a+whole+is+typically+more+significant+(useful,+powerful,+functional,+etc.)+than+the+mere+sum+of+its+parts.&ots=MdYap0L-Cj&sig=Zcq6qtLIpQssrFGsm4GeWhlrHnw

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

Wickens, C. D., Gordon, S. E., Liu, Y., & Lee, J. (2004). *An introduction to human factors engineering* (Vol. 2). Pearson Prentice Hall Upper Saddle River, NJ. https://www.researchgate.net/profile/Christopher-Wickens/publication/239060793_An_Introduction_to_Human_Factors_Engineering/links/00b7d53b8509d91189000000/An-Introduction-to-Human-Factors-Engineering.pdf

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2015). *Engineering psychology and human performance*. Psychology Press.

Wilson, J. R. (2000). Fundamentals of ergonomics in theory and practice. *Applied Ergonomics*, *31*(6), 557–567.

Wilson, J. R. (2014). Fundamentals of systems ergonomics/human factors. *Applied Ergonomics*, *45*(1), 5–13. https://doi.org/10.1016/j.apergo.2013.03.021

Wolstenholme, E. F. (2003). Towards the definition and use of a core set of archetypal structures in system dynamics. *System Dynamics Review*, *19*(1), 7–26. https://doi.org/10.1002/sdr.259

Woods, D., & Dekker, S. (2000). Anticipating the effects of technological change: A new era of dynamics for human factors. *Theoretical Issues in Ergonomics Science*, *1*(3), 272–282.

Wright, R. G. (2020). *Unmanned and autonomous ships: An overview of mass*. Routledge. https://www.taylorfrancis.com/books/mono/10.1201/9780429450655/unmanned-autonomous-ships-glenn-wright

Xie, Z., & Zhong, Z. W. (2016). Aircraft path planning under adverse weather conditions. *MATEC Web of Conferences*, *77*, 15001. https://www.matec-conferences.org/articles/matecconf/abs/2016/40/matecconf_icmmr2016_15001/matecconf_icmmr2016_15001.html

Xu, A., & Dudek, G. (2012). Trust-driven interactive visual navigation for autonomous robots. *2012 IEEE International Conference on Robotics and Automation*, 3922–3929. https://doi.org/10.1109/ICRA.2012.6225171

Xu, A., & Dudek, G. (2015). Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 221–228.

Yamani, Y., & Horrey, W. (2018). A theoretical model of human-automation interaction grounded in resource allocation policy during automated driving. *International Journal of Human Factors and Ergonomics*, *5*, 225. https://doi.org/10.1504/IJHFE.2018.095912

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating Effects of User Experience and System Transparency on Trust in Automation. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, 408–416. https://doi.org/10.1145/2909824.3020230

Yang, X. J., Wickens, C. D., & Hölttä-Otto, K. (2016). How users adjust trust in automation: Contrast effect and hindsight bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 196–200. https://doi.org/10.1177/1541931213601044

Yara. (2018). *Yara Birkeland press kit | Yara International*. Yara None. https://www.yara.com/news-and-media/press-kits/yara-birkeland-press-kit/

Ylikoski, P. (2007). The idea of contrastive explanandum. In *Rethinking explanation* (pp. 27–42). Springer.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

Zausner, T. (2007). Process and Meaning: Nonlinear Dynamics and Psychology in Visual Art. *Nonlinear Dynamics, Psychology, and Life Sciences*, *11*(1), 149.

Ziemke, T. (2008). On the role of emotion in biological and robotic autonomy. *BioSystems*, *91*(2), 401–408.

Zuchowski, L. C. (2018). Complexity as a contrast between dynamics and phenomenology. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, *63*, 86–99. https://doi.org/10.1016/j.shpsb.2017.12.003

Appendix A- Informed Consent Form

Informed Consent Form
"Trust in automation in maritime autonomous surface ships using eye-tracking technology"

Please read this consent agreement carefully before agreeing to participate in this experiment.

This is an inquiry about participation in a research project where the main purpose is to gain an understanding of Trust in Automation using eye-tracking technology in an autonomous ship's simulator. In this letter, we will give you information about the purpose of the project and what your participation will involve.

Purpose of the project
This experimental research is part of a doctoral thesis to build upon the existing knowledge on Trust in Automation in the maritime sector. The project's objective is to develop a dynamic model of trust in automation and to validate the model using the data from an eye-tracking study. In particular, we aim to answer the research question "How does trust in automation change with different degrees of automation reliability?"

Who is responsible for the research project?
University of South-eastern Norway (USN) is the institution responsible for the project.

Why are you being asked to participate?
You are asked to participate in this project because you have familiarity or experience with navigation and/or maritime operations.

What does participation involve for you?
Tobii 2 glasses (Hardware) will be used in this study. The glasses record eye movements, as well as video, and audio of the surroundings on a memory card inserted into the glasses' recorder. While recording, the experimenter can observe the live stream on a laptop which is connected to the glasses. There will be no recordings of the events but only live transmission from the glasses. Participants will be identified as "Participant 1", "Participant 2", and so on.
All recordings will then be transferred to a USN's computers and stored in a secured drive, processed with Tobii Pro software for initial screening and visualizations, and then exported as CSV files for further statistical analysis in R. Tobii Pro as the software will be used for data analysis. The software is installed locally on a USN's laptop and does not sync up with cloud storage.

At the beginning of the study, you will be familiarized with the overall scope of the study, tools, and equipment used in this study. Next, you will be asked to read, understand, and sign the consent form. Upon the agreement to participate in the experiment, you will be requested to fill out two paper-based surveys including a) Demographic information regarding your age, gender, education, and experience, and b) A personality test (Big Five).

The experiment will last approximately 35 minutes. During the experiment, you will be asked to wear eye-tracking glasses (Tobii 2 glasses) and perform a supervisory control of unmanned vessels (detailed instructions will be provided). Data regarding the eye movements (gaze, fixation, and saccades) as well as audio and video recordings of the environment will be collected. At the middle and end of the study, you are asked to fill out a subjective assessment form regarding your level trust in the automated system via paper/electronically recorded surveys.

Participation is voluntary
Participation in the project is voluntary. If you chose to participate, you can withdraw your consent at any time without giving a reason. All information about you will then be made anonymous. There will be no negative consequences for you if you chose not to participate or later decide to withdraw.

Your personal privacy – how we will store and use your personal data
We will only use your personal data for the purpose(s) specified in this information letter. We will process your personal data confidentially and in accordance with data protection legislation (the General Data Protection Regulation and Personal Data Act). Only the project team and the data analyst will have access to the personal data. The personal data will be secured within the course of this study and be coded/encrypted after the analysis. Participants will not be recognizable in publications of this study.

What will happen to your personal data at the end of the research project?
The project is scheduled to end on 15.11.2023. The personal data, including any digital recordings, will be anonymized at the end of the project.

Your rights
As long as you can be identified in the collected data, you have the right to:

- access the personal data that is being processed about you
- request that your personal data is deleted
- request that incorrect personal data about you is corrected/rectified
- receive a copy of your personal data (data portability), and
- send a complaint to the Data Protection Officer or The Norwegian Data Protection Authority regarding the processing of your personal data

What gives us the right to process your personal data?
We will process your personal data based on your consent.
Based on an agreement with University of South-eastern Norway (USN), Data Protection Services has assessed that the processing of personal data in this project is in accordance with data protection legislation.

Where can I find out more?
If you have questions about the project or want to exercise your rights, contact:

- University of South-eastern Norway (USN) via Mehdi Poornikoo mehdi.poornikoo@usn.no
- Our Data Protection Officer: Paal Are Solberg paal.a.solberg@usn.no
- Data Protection Services, by email: (personverntjenester@sikt.no) or by telephone: +47 53 21 15 00.

Yours sincerely,
Mehdi Poornikoo

Consent form

I have received and understood information about the project "Trust in automation in maritime autonomous surface ships using eye-tracking technology" and have been given the opportunity to ask questions. I give consent:

- ☐ to participate in the eye-tracking experimental study
- ☐ to participate in paper/electronic surveys
- ☐ my personal data to be processed until the end date of the project, approx. *15.11.2023*

Signature: _____

Date: _____

Name: _____

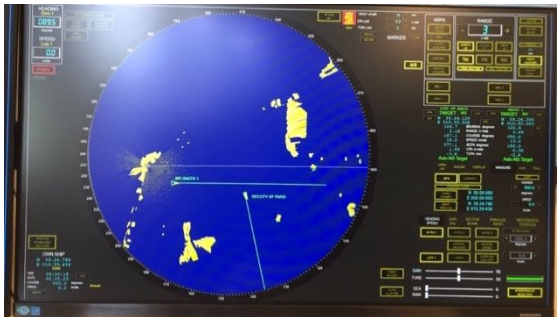
Participant ID: _____

Appendix B- Experiment Instruction

Instruction

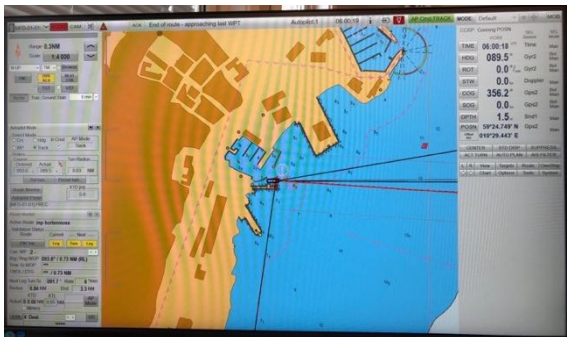
Checklist Setup:

1. **RADAR:**
   - INTERFERENCE REJECTION: ON
   - NAVIGATE: GPS: ON
   - AIS: ON
   - RANGE RINGS: OFF
   - RANGE: 3 miles
   - RM: ON



2. **Main Interface:**
   - SPEED: ACTIVE: ON,
   - INTERLOCK: ON,
   - Speed: %100 ~ 13 Kn

3. **ECDIS:**
   - Route: Manage Route: mp hortenmoss, Monitor
   - Route: Route Monitor |<
   - Current WP: 02
   - AP Mode: ON |<
   - Autopilot: Track
   - Themes: Non-chart: past track to remove previous tracks



Briefing

Vessel Info:

Small Ro-Ro carrier
The vessel has 2 propellers & 2 rudders

Route:

Horten-Moss on pre-validated route

Weather:

Wind South 6 Kn.
Significant waves: 0.4

Traffic:

Bastø 1: Eastbound,
Bastø 2: Westbound

Responsibility:

Only sailing
The mooring crew will take over

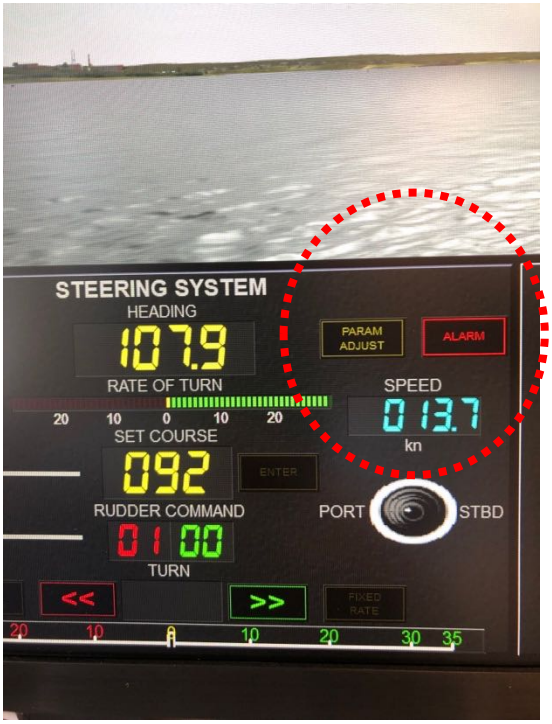
The Vessel starts in Autopilot & Autotrack mode with a speed of 100%.

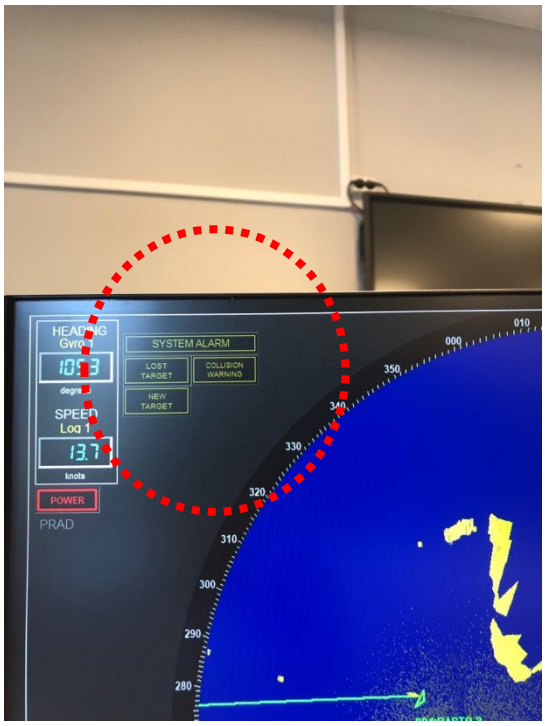
*Please do not change the ECDIS layout

Your tasks:

1. Supervise the vessel's performance for a safe operation
2. Ensure the vessel follows the pre-validated route
3. In case of error or deviation:

   ☐ Acknowledge the error by clicking on the **ALARM buttons**, located on the top left corner of the RADAR screen, and the STEERING SYSTEM, as shown below:





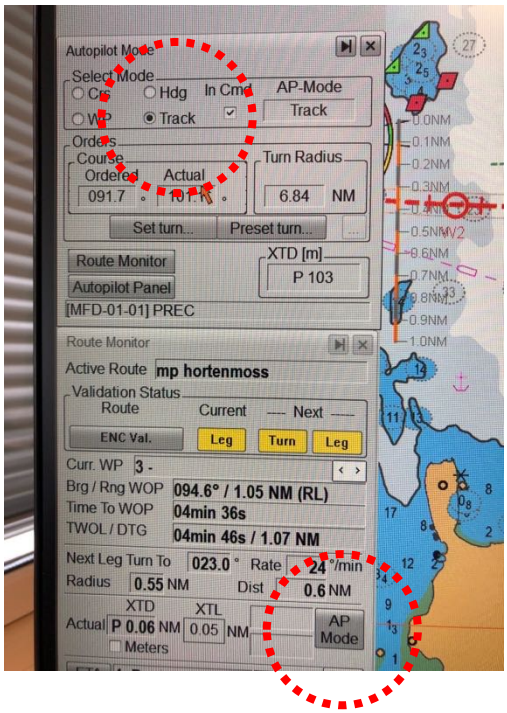
☐ **Set the vessel on track** by choosing either:

- **Manual** control using TURN Slider for RUDDER COMMAND

- **Autopilot**: SET COURSE to the Next WP

- Re-activate **Nav Auto track** by selecting NAV in the STEERING SYSTEM, Selecting AP Mode, and Track command in ECDIS


STEERING SYSTEM
HEADING
093.6
PARAM ADJUST
NAV
RATE OF TURN
20 10 0 10 20
SPEED
012.1
kn
SET COURSE
AUTO
094
ENTER
RUDDER COMMAND
PORT
STBD
MANUAL
00 00
TURN
NFU
FIXED RADIUS
<<
>>
FIXED RATE
35 30 20 10 0 10 20 30 35


Autopilot Mode
Select Mode
Crs
Hdg
In Cmd
AP-Mode
WP
Track
Track
Orders
Course
Ordered
Actual
Turn Radius
091.7 °
6.84 NM
Set turn...
Preset turn...
Route Monitor
XTD [m]
Autopilot Panel
P 103
[MFD-01-01] PREC
Route Monitor
Active Route
mp hortenmoss
Validation Status
Route
Current
Next
ENC Val.
Leg
Turn
Leg
Curr. WP
3 -
Brg / Rng WOP
094.6° / 1.05 NM (RL)
Time To WOP
04min 36s
TWOL / DTG
04min 46s / 1.07 NM
Next Leg Turn To
023.0 °
Rate
24 °/min
Radius
0.55 NM
Dist
0.6 NM
XTD
XTL
Actual
P 0.06 NM
0.05 NM
AP Mode
Meters

Appendix C- Demographic Form

"Trust in automation in maritime autonomous surface ships using eye-tracking technology".

Demographic information

Participant No.: _____

Age:

- ☐ 18-24
- ☐ 25-34
- ☐ 35-44
- ☐ 45-54
- ☐ 55-64

Gender:

- ☐ Male
- ☐ Female
- ☐ Prefer not to say

Education:

- ☐ High school degree
- ☐ Bachelor's degree (e.g., BA, BS)
- ☐ Master's degree (e.g., MA, MSc.)
- ☐ Doctorate (Ph.D.)

Seafaring experience:

- ☐ None
- ☐ 1-5 years
- ☐ 5-10 years
- ☐ 10-15 years
- ☐ Over 15 years

Knowledge of Navigation system

- ☐ Limited
- ☐ Some
- ☐ Good
- ☐ Very good
- ☐ Expert

Appendix D- Trust in Automation Questionnaire

		Strongly disagree	Rather disagree	Neither disagree nor agree	Rather agree	Strongly agree	No response
1	The system is capable of interpreting situations correctly.	(1)	(2)	(3)	(4)	(5)	○
2	The system state was always clear to me.	(1)	(2)	(3)	(4)	(5)	○
3	I already know similar systems.	(1)	(2)	(3)	(4)	(5)	○
4	The developers are trustworthy.	(1)	(2)	(3)	(4)	(5)	○
5	One should be careful with unfamiliar automated systems.	(1)	(2)	(3)	(4)	(5)	○
6	The system works reliably.	(1)	(2)	(3)	(4)	(5)	○
7	The system reacts unpredictably.	(1)	(2)	(3)	(4)	(5)	○
8	The developers take my well-being seriously.	(1)	(2)	(3)	(4)	(5)	○
9	I trust the system.	(1)	(2)	(3)	(4)	(5)	○
10	A system malfunction is likely.	(1)	(2)	(3)	(4)	(5)	○
11	I was able to understand why things happened.	(1)	(2)	(3)	(4)	(5)	○
12	I rather trust a system than I mistrust it.	(1)	(2)	(3)	(4)	(5)	○
13	The system is capable of taking over complicated tasks.	(1)	(2)	(3)	(4)	(5)	○
14	I can rely on the system.	(1)	(2)	(3)	(4)	(5)	○
15	The system might make sporadic errors.	(1)	(2)	(3)	(4)	(5)	○
16	It is difficult to identify what the system will do next.	(1)	(2)	(3)	(4)	(5)	○
17	I have already used similar systems.	(1)	(2)	(3)	(4)	(5)	○
18	Automated systems generally work well.	(1)	(2)	(3)	(4)	(5)	○
19	I am confident about the system's capabilities.	(1)	(2)	(3)	(4)	(5)	○

Questionnaire „Trust in Automation“ (TiA) | Moritz Körber, Technical University of Munich

Based on your observation, please answer the following questions:

1. Did you notice any errors?

☐ YES
☐ NO

2. What do you think could be the source of the error?

……

Appendix E- The Big Five Personality Test Questionnaire

The Big Five Personality Test
from personality-testing.info
courtesy ipip.ori.org

Introduction

This is a personality test, it will help you understand why you act the way that you do and how your personality is structured. Please follow the instructions below, scoring and results are on the next page.

Instructions

In the table below, for each statement 1-50 mark how much you agree with on the scale 1-5, where 1=disagree, 2=slightly disagree, 3=neutral, 4=slightly agree and 5=agree, in the box to the left of it.

Test

Rating	I....	Rating	I.....
	1. Am the life of the party.		26. Have little to say.
	2. Feel little concern for others.		27. Have a soft heart.
	3. Am always prepared.		28. Often forget to put things back in their proper place.
	4. Get stressed out easily.		29. Get upset easily.
	5. Have a rich vocabulary.		30. Do not have a good imagination.
	6. Don't talk a lot.		31. Talk to a lot of different people at parties.
	7. Am interested in people.		32. Am not really interested in others.
	8. Leave my belongings around.		33. Like order.
	9. Am relaxed most of the time.		34. Change my mood a lot.
	10. Have difficulty understanding abstract ideas.		35. Am quick to understand things.
	11. Feel comfortable around people.		36. Don't like to draw attention to myself.
	12. Insult people.		37. Take time out for others.
	13. Pay attention to details.		38. Shirk my duties.
	14. Worry about things.		39. Have frequent mood swings.
	15. Have a vivid imagination.		40. Use difficult words.
	16. Keep in the background.		41. Don't mind being the center of attention.
	17. Sympathize with others' feelings.		42. Feel others' emotions.
	18. Make a mess of things.		43. Follow a schedule.
	19. Seldom feel blue.		44. Get irritated easily.
	20. Am not interested in abstract ideas.		45. Spend time reflecting on things.
	21. Start conversations.		46. Am quiet around strangers.
	22. Am not interested in other people's problems.		47. Make people feel at ease.
	23. Get chores done right away.		48. Am exacting in my work.
	24. Am easily disturbed.		49. Often feel blue.
	25. Have excellent ideas.		50. Am full of ideas.

Appendix F- STELLA Syntax Documentation

	Top-Level Model:	Equation	Properties
	Expectation_of_performance(t)	Expectation_of_performance(t - dt) + (change_in_expectation) * dt	INIT Expectation_of_performance = INITIAL_EXPECTATION
	Perceived_performance(t)	Perceived_performance(t - dt) + (change_in_perceived_performance) * dt	INIT Perceived_performance = INITIAL_PERCEIVED_PERFORMANCE
	Trust(t)	Trust(t - dt) + (change_in_TiA) * dt	INIT Trust = INITIAL_TRUST
	change_in_expectation	((Trust-Expectation_of_performance)/adjustment_time_expectation)+Expectation_ramp	
	change_in_perceived_performance	System_performance*fractional_perception_rate*Perceived_performance*(Maximum_capability-Perceived_performance)/ad_time_performance	
	change_in_TiA	MIN(Expectation_gap/adjustment_time_TiA; Discrepancy/adjustment_time_TiA)	
	ad_time_performance	3	
	adjustment_time_expectation	3	
	adjustment_time_TiA	Baseline_AD_time/Propensity_Factor	
	Baseline_AD_time	2,5	
	Baseline_challenge	0	
	Desire_to_allocate_trust_behavior	0,1-Perceived_risk_1	
	Discrepancy	Max_TiA_faith-Trust	
	Environmental_challenges	Baseline_challenge+(SWITCH_STEP_1_ON*STEP_Challenge)	
	Error_magnitude	50	
	Error_Time	25	
	Expectation_gap	Perceived_performance-Expectation_of_performance	
	Expectation_ramp	RAMP(Expectation_slope; 50; 60)	
	Expectation_slope	0	
	fractional_perception_rate	0,0005	
	INITIAL_EXPECTATION	30	
	INITIAL_PERCEIVED_PERFORMANCE	30	
	INITIAL_TRUST	30	

○	Max_TiA_faith	100	
○	Maximum_capability	Nominal_capability-Environmental_challenges	
○	Nominal_capability	100	
○	Perceived_risk_1	0	
○	Propensity_Factor	0,7	
○	STEP_Challenge	STEP(30; 50; 10; 0)+STEP(-30; 75; 10; 0)	
○	SWITCH_1_ON	0	
○	SWITCH_STEP_1_ON	0	
○	System_Malfunction	SWITCH_1_ON*PULSE(-Error_magnitude; Error_Time; 0)	
○	System_performance	"Trusting_behavior,_Use"+System_Malfunction	
○	"Trusting_behavior,_Use"	(Trust*Desire_to_allocate_trust_behavior)	

Run Specs	
Start Time	0
Stop Time	100
DT	1/125
Fractional DT	True
Save Interval	0,008
Sim Duration	1,5
Time Units	Time
Pause Interval	0
Integration Method	Euler
Keep all variable results	True
Run By	Run
Calculate loop dominance information	True
Exhaustive Search Threshold	1000

Appendix G- Articles

Article 1

Poornikoo, M., & Øvergård, K. I. (2023). Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation. Theoretical Issues in Ergonomics Science, 1-37. https://doi.org/10.1080/1463922X.2023.2233591

Article 2

Poornikoo M., Mansouri M. (2023), "Systems approach to modeling controversy in Human factors and ergonomics (HFE)," 18th Annual System of Systems Engineering Conference (SoSe), Lille, France, 2023, pp. 1-8,
https://doi.org/10.1109/SoSE59841.2023.10178634

Article 3

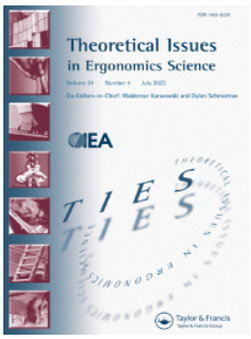
Poornikoo, M., & Øvergård, K. I. (2022). Levels of automation in maritime autonomous surface ships (MASS): A fuzzy logic approach. Maritime Economics & Logistics, 24(2), 278-301. https://doi.org/10.1057/s41278-022-00215-z

Article 4

Poornikoo M., Gyldensten W., Vesin B., Øvergård, K. I. (In review) Trust in Automation (TiA): simulation model, and empirical findings in supervisory control of Maritime Autonomous Surface Ships (MASS), International Journal of Human-Computer Interaction

Article 1

Poornikoo, M., & Øvergård, K. I. (2023). Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation. Theoretical Issues in Ergonomics Science, 1-37. https://doi.org/10.1080/1463922X.2023.2233591



Theoretical Issues in Ergonomics Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttie20

Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation

Mehdi Poornikoo & Kjell Ivar Øvergård

To cite this article: Mehdi Poornikoo & Kjell Ivar Øvergård (2023): Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation, Theoretical Issues in Ergonomics Science, DOI: 10.1080/1463922X.2023.2233591

To link to this article: https://doi.org/10.1080/1463922X.2023.2233591

Published online: 13 Jul 2023.

THEORETICAL ISSUES IN ERGONOMICS SCIENCE
https://doi.org/10.1080/1463922X.2023.2233591

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS

Check for updates

Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation

Mehdi Poornikoo[a] and Kjell Ivar Øvergård[b]

[a]Department of Maritime Operations, University of South-Eastern Norway (USN), Horten, Norway; [b]Department of Health, Social and Welfare Studies, University of South-Eastern Norway (USN), Horten, Norway

ABSTRACT

Theories and models are central to Human Factors/Ergonomics (HFE) sciences for producing new knowledge, pushing the boundaries of the field, and providing a basis for designing systems that can improve human performance. Despite the key role, there has been less attention to what constitutes a good theory/model and how to examine the relative worth of different theories/models. This study aims to bridge this gap by (1) proposing a set of criteria for evaluating models in HFE, (2) employing a methodological approach to utilize the proposed criteria, and (3) evaluating the existing models of trust in automation (TiA) according to the proposed criteria. The resulting work provides a reference guide for researchers to examine the existing models' performance and to make meaningful comparisons between TiA models. The results also shed light on the differences among TiA models in satisfying the criteria. While conceptual models offer valuable insights into identifying the causal factors, their limitation in operationalization poses a major challenge in terms of testability and empirical validity. On the other hand, although more readily testable and possessing higher predictive power, computational models are confined to capturing only partial causal factors and have reduced explanatory power capacity. The study concludes with recommendations that in order to advance as a scientific discipline, HFE should adopt modelling approaches that can help us understand the complexities of human performance in dynamic sociotechnical systems.

ARTICLE HISTORY
Received 26 December 2022
Accepted 2 July 2023

KEYWORDS
Scientific criteria; model evaluation; trust in automation; folk models; modelling approach

Relevance to human factors/ergonomics theory

For human factors and ergonomics (HFE) as a discipline to progress, it is necessary to produce and validate scientific theories and models. Testing and evaluating models are essential aspects of the theory/model development process, allowing for the recognition of advancements in the field. This study proposes a number of criteria for model evaluation in HFE and a methodological procedure to apply these criteria to the models of trust in automation.

Introduction

A long-standing discussion in Human Factors/Ergonomics (HFE) is whether constructs and models are 'folk models'; that is, whether they are credible and scientific (Dekker and Hollnagel

CONTACT Mehdi Poornikoo mpo@usn.no

2004; Flach 1995; Sarter and Woods 1991; van Winsen and Dekker 2015). The term 'folk psychology' is referred to the 'collection of psychological principles and generalizations which, … underlies our everyday explanation of behaviour' (Stich and Nichols 1992, 37). People can make remarkably well-articulated naïve theories of motion based on their everyday experiences. Such theories are sensible outcomes of interactions with the real world, which may not be consistent with the principles of physics but tend to continue as a common-sense and laypeople's explanation of the physical world (McCloskey 1983). Similarly, psychology has been populated with folk models of human behaviour, which are not necessarily wrong but compared to more articulated models, they tend to focus on descriptions rather than explaining phenomena, making them very hard to test and falsify (Corbett 2015).

Within the HFE discipline, Dekker and Hollnagel (2004) have raised concerns regarding the scientific credibility of several theoretical constructs (e.g. situation awareness and trust in automation) and their relation to human performance. Several researchers have presented claims that these constructs are theoretically unclear, unfalsifiable, excessively generalizable, and with generic descriptive labels rather than proper explanations for causal psychological mechanisms relevant to the performance (Cass 2011; Douglas, Aleva, and Havig 2007; Flach 1995; Jodlowski 2008). Billings claims that HFE constructs have become too neat and too holistic (Billings 1995) relying on their face validity as intuitive concepts (Jones 2015). Yet, face validity is considered the weakest form of validity (Drost 2011).

In opposition to Dekker and Hollnagel (2004) some scholars (e.g. Endsley 2015; Parasuraman, Sheridan, and Wickens 2008; Wickens 2008) argue that a large body of research on situation awareness, mental workload, and trust in automation (TiA) indicates the credibility of these constructs and their practical usefulness. Parasuraman, Sheridan, and Wickens (2008) maintain that Popper's (1972) notion of falsification has less relevance for theory development in cognitive engineering and ergonomics sciences because these constructs are not part of empirical reality or statement of fact and therefore, falsifiability of such constructs becomes a meaningless idea. According to Parasuraman, Sheridan, and Wickens (2008), HFE constructs are scientifically credible and should not be held accountable for being proven as 'right or wrong' but instead, attempts should be directed to 'establish contextual limitations in which a theory or principle successfully predicts performance and makes testable recommendations…' (Parasuraman, Sheridan, and Wickens 2008, 155).

The divergent perspectives on the credibility of HFE constructs call for a critical review of the existing theories and models in the HFE discipline. We believe a viable solution to the folk-model controversy is not to take a general 'yes or no' position but rather to promote a framework that will allow us to assess the scientific nature of theories by examining their epistemological assumptions, quality of propositions, and empirical adequacy. The purpose of this paper is then twofold. The first section sets forth a set of criteria for evaluating scientific theories in HFE, which can lead to cumulative scientific progress in the field. In the second part of the study, we review the existing Trust in Automation (TiA) models (which is one of the theories being accused of being a Folk Model construct) and probe these models against the proposed criteria to compare the efficacy of the models for real-world use. By doing so we hope to be able to assess whether the TiA research programme (Lakatos 1978) is progressing or not.

Theory evaluation in HFE

One of the general aims of science is to produce and test theories (Kerlinger and Lee 1986). Theories are central to scientific understanding because they allow us to see relationships

between phenomena that might otherwise appear disconnected. Theories also illuminate the underlying causes or structure of a phenomenon and thus enable us to develop successful interventions to consolidate or prevent a particular effect (Risjord 2019).

Underlying any form of scientific inquiry is a philosophy of science that elucidates a researcher's approach to the nature of the phenomenon being studied (ontology) and the methods for comprehending it (epistemology). Whether explicitly or implicitly, we rely on the philosophy of science to understand the meanings, logical relationships, and consequences of our theoretical assertions and observations (Van de Ven 2007). Philosophers have endlessly debated these topics and developed a variety of research philosophies for what constitutes science and scientific progress. In a realistic view of science (Scientific Realism), the progress of science is furthered by empirical testing of theories that allow theories to encompass more and more of empirical phenomena, thereby improving the 'truthlikeness' of the theory (Niiniluoto 1999). Not all agree with this goal for science, and some view science as a problem-solving activity where scientific progress is achieved when theories can help solve new problems (e.g. Azevedo 1997; Campbell 1988; Deutch 1998; Laudan 1978). Irrespective of the nuances of this long-standing disagreement between these two views of philosophy of science, truthlikeness and problem-solving ability are not mutually exclusive goals. A theory becomes useful the moment it describes and can predict how part(s) of the world work. A theory that has no relation to how the world works can only spuriously hope to improve the solution of problems as the use of the theory would actually be based upon wrong presuppositions, and if so – the problem-solving element of the theory would be pure luck – based upon coincidences and not upon a thorough understanding of how the world works. On the other hand, a theory that has truthlikeness and encompasses and explains observed data would probably be of more practical value than a theory that does not explain the observed data. Likewise, a theory that improves the problem-solving activity in the physical world probably also has a higher truthlikeness. Hence, we would claim – in accordance with Niiniluoto (2017) – that there is a correspondence between truthlikeness and problem-solving ability, thus pointing out that the practical consequences of the realistic- and pragmatic orientations to science are similar – theories allow us to understand, explain, and act on the world in order to do new things.

The likelihood that a theory will be rejected determines how credible the theory is (Van de Ven 2007). According to Popper (1972), a theory must be falsifiable or otherwise deemed as a pseudo-scientific theory. Although the idea of falsification by a single study (what Lakatos (1978) has called naïve falsification) has been met with heavy critique (e.g. Kuhn 1962) and subsequently refined by pointing out that falsification requires multiple refutations and the presence of an alternative and superior theory (Lakatos 1978), the idea of theory evaluation is a cornerstone of the scientific methods (Carnap 1953; Lakatos 1970, 1978; Popper 1972; Ngwenyama 2014). Scientists collect and report data to test and evaluate theories (Trafimow 2012), yet it is not easy to think of theories in social sciences and psychology that are clearly falsified (Van Lange 2013). Whether one prefers hard falsification (Popper 1972), a softer version of falsification (Lakatos 1978), strong inference falsification (Platt 1964), or Bayesian inference (Edwards, Lindman, and Savage 1963; Howson and Urbach 1989), theories must be empirically testable (falsifiable) and closely correspond to the investigated phenomenon.

That said, empirical testability cannot be a single criterion as an unclear theory is able to accommodate any observation consistent with itself (Deutsch 2011). As Lakatos (1970, 184) puts it: 'Any theory … can be saved from refutation by some suitable adjustment in

the background knowledge'. Therefore, falsifiability and empirical evidence are necessary conditions but not sufficient criteria for assessing the credibility of a theory or at least the relative worth of alternative theories. Van de Ven (2007) advocates that theories cannot be justified only by testing their empirical fit with the real world but rather by rhetorical arguments about the logical validity of a theory. A good theory is expected to offer clear operational definitions, internal logical consistency, verifiability (Bacharach 1989; Péli and Masuch 1997; Wacker 2004), and replicability of findings that are obtained from a precisely-stated theory (Earp and Trafimow 2015).

Middle-range theories as models

Theories consist of constructs (abstract ideas or concepts) that are connected in a logical way (Baumeister and Bushman 2020), which is defined as 'a set of abstract concepts (i.e. constructs) together with propositions about how those constructs are related to one another' (Manstead and Livingstone 2008, 27). Theories are usually not open to direct examination, while models can make specific predictions of theory that can be tested (Van de Ven 2007). The high level of abstraction in theories often resists falsification (Weick 1974).

Models typically consist of symbols that specify the characteristics of a phenomenon, its components, and relationships among the components. Though there is no well-defined distinction between theories and models, a theory appears like a narrative description, while a model can be analogous to a map. Models enable researchers to formulate empirically testable propositions about aspects of a theory (Frankfort-Nachmias, Nachmias, and DeWaard 2014) and hence can be regarded as partial representations of theories. The empirical investigation is commonly achieved *via* modelling. Social scientists do not directly observe and test theories; instead, they study and inspect models (McKelvey 2017). Models may also encompass procedures, assumptions, and manipulations that are used to apply the scientific methodology of observation and analysis. These assumptions and procedures are not typically embedded in the theory itself; therefore, a model is not just an operational version of a theory but rather acts as a mediator or intermediary between theory and empirical evidence (Morgan and Morrison 1999).

Theories can be classified based on their level of abstraction. Merton (1968) provides a distinction between 'grand' and 'middle-range' theories. Grand theories are the most abstract, normative, unbounded, and all-encompassing theories that address the nature, mission, and purpose of a phenomenon in a fairly general fashion (Peterson and Bredow 2013). Compared to grand theories, middle-range theories are less abstract, narrower in scope and specificity, and more readily usable and testable in research projects. In other words, middle-range theories are abstract enough to allow for generalizations but specific enough for observed data to be incorporated into propositions that can be empirically tested. Based on this categorization, one can think of HFE's theories as middle-range theories, also frequently referred to as theoretical constructs. Theoretical constructs are invented terms that can neither be directly nor indirectly observed but may be entirely defined based on observable variables (Kaplan 1964).

Risjord (2019) argues that middle-range theories can be better understood when analysed as models. We usually differentiate theories by referring to specific models. This is particularly relevant in HFE studies as, for example, theories of trust in automation (TiA) are commonly discussed as Muir's (1994) integrated model of trust in human-machine

relationships or Lee and See (2004) conceptual model of trust and reliance. By focusing on models, we shift our attention from the structure to the core content of the theory. Models emphasize causality and demonstrate how some events occur because of processes and interactions among the model elements.

Causal relationships in models can help HFE professionals to identify potential areas for improving human performance in sociotechnical systems. Furthermore, considering HFE theories as models forges a stronger link between the adequacy of the model and the motivations/occasions for using them. That is, since models are analogous to maps, they ignore some aspects of reality to be simple and useful. A street map of Paris creates an abstraction of the world – ignoring many aspects not directly relevant to navigation – to simplify navigation through the streets of Paris. Different models then represent different features of the same thing for different purposes. It means a model implicitly assumes some features to be more important than others. This is why multiple models based on different assumptions and background theories are often needed to comprehend complicated phenomena (Fried 2020; Risjord 2019). Lastly, models specify interactions and allow us to test whether changes in one element's activity can change the others, as explained by the model. We then evaluate the model's empirical support and highlight its accuracy for applications in real-world settings. Model evaluation focuses on the phenomenon being modelled, its fundamental assumptions, the elements of the model, and the relationships between its elements (Degani and Heymann 2002).

It is also important to distinguish between theoretical models from statistical models. While theoretical models represent phenomena in the world and propose global conjectures about aspects of a phenomenon, statistical models are data models that represent data and are used for testing hypotheses locally, derived from theory and through the process of hypothetico-deductive framework (Borsboom et al. 2021; Robinaugh et al. 2021). Despite close correspondence, theoretical and statistical models should not be confused. The former deals with scientific epistemology and justification of knowledge, while the latter involves scientific methodology and justification of methods (Carter and Little 2007). Although questions about methodology are beyond the scope of this study, a review of empirical findings and statistical methods is necessary to investigate the empirical adequacy of the existing models.

Criteria development to evaluate HFE models

Theory evaluation is not possible without a set of criteria by which it is to be evaluated. The challenging parts of theory evaluation, however, are the appropriateness and use of epistemological criteria for evaluating theories (Howard 1985). While providing a list of criteria seems rather easy, scholars may disagree on how to apply these criteria, their relative significance, and the degree to which a theory/model is supported by a given criterion. Laudan (1986) reminds us that theoretical disagreements may happen at any level (substantive, content, or methodological levels), which are to some extent subject to the aim of science. Unfortunately, epistemological criteria cannot tell us what the aim of science - especially in social sciences - should be (Witkin and Gottschalk 1988). The choice of criteria for theory evaluation is ultimately dependent on the evaluator's view on ontology, epistemology, methodology, and purpose (Prochaska, Wright, and Velicer 2008).

To develop a set of criteria for evaluating HFE models, a review of leading philosophers of science (e.g. Blalock 1969; Dubin 1970; Kuhn 1977; Meleis 2012; Popper 1969; Van de Ven, 2007), combined with Kivunja's (2018) systematic literature review on the fundamental constituents of a scientific theory is performed. While most of these criteria are widely established principles for theory assessment, some are specific to the phenomenon under investigation (Here, TiA).

Criterion 1: testability/falsifiability

Testability or falsifiability (Popper 1969) is an essential part of science and is often regarded as the most rigorous criterion (Cramer 2013). If a model is not testable, we cannot assess its empirical value. Testability is typically considered an empirically-based criterion. While the relatively abstract and general nature of grand theories may hinder direct measurement and operationalization of the concepts, the relatively concrete and precise nature of middle-range theories means that they can have operational definitions, and their propositions must be open to direct empirical testing (Saunders, Lewis, and Thornhill 2007).

To assess the testability of the middle-range theories (i.e. HFE models) a classical empiricism approach would demand that the concepts of the theory are observable, and the propositions are quantifiable (Fawcett 2005). Concepts would be empirically observable when operational definitions provide empirical indicators that are used to identify the concepts. Propositions then can be examined when empirical indicators can be replaced with the concepts and when methods can adequately give proof for the assertions made (Fawcett 1988). A substantial advantage of representing HFE middle-range theories as models is that it highlights the ways that the models can be tested. If the chosen model is operationalized and relatively precise, the relevant test can signify whether the model's components change in the way that the model predicts. Such tests are direct tests of the model and indicate the relationship between the construct of interest and the empirical observations.

Although nonempirical tests such as computer simulation can be beneficial when contextual details are well-incorporated in the model, often it is the empirical research that can give support (or lack of it) to the model. At the operational level, testability has also important implications for the methods that are available. For instance, recent developments in neuroscience and its techniques, such as fMRI, allow researchers to test assertions that previously could not be possible. When evaluating the testability of HFE models, we adapt Fawcett (1986, 2005) and Silva (1986) three main questions:

(1) Can the model be operationalized? Is there a way of measuring the components and constructs in the model?
(2) Does the model suggest a research design for testing its assumptions?
(3) Are the measurement tools and data analysis techniques adequate to measure the model propositions?

Criterion 2: predictive power

To employ the testability criterion, a model/theory must make some predictions. According to Popper (1969), the more specific predictions one can make, the better it is,

as specific predictions are riskier and therefore more likely to fail, and hence it is easier to falsify the theory. For example, a linear relationship between two variables stated as 'A is correlated with B' rules out practically nothing except when the correlation is zero, while 'A is positively correlated with B' makes a more specific prediction by ruling out 50% of possible outcomes. The latter statement is more falsifiable and would constitute a better form of theory than the former. Thus, a model is better the more precise predictions it makes. As long as there is a pathway in a causal model which is testable, the model potentially has a degree of predictive power (Dienes 2008). Meehl (1978) points out a difference between point prediction (predicting a particular parameter value) and directional prediction (predicting the direction of an effect – e.g. positive or negative). Point prediction is typically common for 'harder' sciences such as physics and chemistry, which indicates the rigor of precision. This precision has been attributed to the neatly interrelated and tightly connected components and constructs in physical sciences. Theories in social sciences and psychology, on the other hand, tend to focus on directional prediction.

Prochaska, Wright, and Velicer (2008) promote predictions of effect sizes between constructs in order for theories to provide riskier predictions. Effect size estimates make tighter and more explicit quantitative predictions. This would also help researchers to go beyond pure reliance on null hypothesis testing and its limitations for the theory evaluation (Prochaska, Wright, and Velicer 2008). That said, we advocate a differentiation of quantitative predictions in HFE models according to a simple-to-complex listing of predicted empirical/causal relations. The models that make the more complex predictions are deemed to have a higher scientific level (given that the model's predictions are correct). The criteria for determining the scientific level of a model´s predictions are described from 'simple' to 'complex' below.

(1) **Predicting the Existence of an effect**: Specifying the existence or non-existence of a relationship between constructs. In a path model, this would be akin to adding or removing an arrow connecting two constructs (Pearl 2009). This is the simplest prediction and is similar to the standard null hypothesis test.
(2) **Predicting the direction (or sign) of an effect**: Specifying the direction of effects – e.g. construct A is positively correlated with construct B.
(3) **Predicting the size and direction of the effect**: Specifying the direction and size of the effect – e.g. constructs A and B will have a correlation $r = 0.40$. Even better would be adding a prediction for the variance of the observed effect. This could be shown by presenting a Confidence Interval (CI) for the effect.
(4) **Mathematical specification of the form of the predicted effect:** Another improvement on points 1–3 is the specification of the mathematical form of relationships between variables. This is often forgotten in psychology as most mathematical/statistical predictions use an assumption of linearity (Freedman 2010; McElreath 2018); however, we know that many (if not most) relationships are non-linear in nature (Guastello 2001, 2017; Thompson, Stewart, and Turner 1990). Hence, specifying not only the direction and size of a relationship but also the mathematical form of a relationship – so that we know if a relationship is assumed to be linear (e.g. $y = a + bx$), curvilinear (e.g. $y = a + bx + bx^2$) or non-linear (e.g. $y = ax^2 + bx^3$) – would improve the testability of a theory.

These four sub-criteria are directionally complimentary as any model whose predictions fulfil sub-criterion 3 will automatically also fulfil sub-criteria 1 and 2, while a model that only fulfils sub-criterion 1 will not satisfy sub-criteria 2–4.

Criterion 3: explanatory power

One problem of incomplete theories is that they often make some predictions but are unable to provide an adequate explanation of the phenomenon. Ancient astronomers were able to make accurate predictions without satisfactory explanation (Kaplan 1964). A model is useful when it can both predict and explain (Bacharach 1989). Indeed, prediction and explanation are two sides of the same coin and complementary characteristics of a good theory. Explanations that implore causal relationships always make predictions, particularly predictions on future events under causal intervention. Even if predictions are not declared explicitly, the language of causal explanation often implies a sequence of events as the 'reason' for some specific outcomes (Hofman, Sharma, and Watts 2017).

Cramer (2013) exemplifies explanatory power in the process of reckoning the next value in a series of numbers as 1 2 3 5 8 … Since there can be different ways to predict the next number by adding and subtracting various combinations, explanation provides logic and justification for the predicted outcome. Theories should therefore have a priori truthlikeness or verisimilitude; i.e. they must be viable and produce explananda before testing (Fried 2020). 'One needs theory first to know what is worth testing' (Van Rooij and Baggio 2021, 324). This criterion is greatly applicable to applied problems in the HFE domain. Applied problems require an understanding of the phenomenon by virtue of a complete explanation and particular predictions of the outcome (Athey 2017).

Appropriate explanations in science necessitate clear proof of causality (Prochaska, Wright, and Velicer 2008). One approach is to create experimental control, which is normally accomplished using an experiment where you control the presence of independent variables and measure the changes in a dependent variable. The changes in the dependent variable can then be explained by the manipulation of the independent variable. However, in real-world contexts, experimental control is often not possible or is very hard to achieve, and this is particularly so for behaviours and phenomena that are critical to the HFE field.

Statistical control is an alternative when experimental control is not feasible or ethical to use. With statistical control, the association between an independent and a dependent variable is controlled for by removing the variation explained by other independent variables, like in a multiple regression model (Cohen et al. 1983). Theoretical models, controlled experiments, and statistical control are all means to acquire causal knowledge by inquiring about how changes in a set of causal factors change the outcome (Woodward 2005). Since different models may portray different causal factors for a particular phenomenon, the causal explanation can be regarded as 'interest relative' (Lipton 1990). This implies that a model should elucidate not only 'why this' but 'why this rather than that' for a set of causal factors. This view fits with the contrastive account of explanation (Garfinkel 1982; Lipton 1990; Ylikoski 2007), which demonstrates how models are used to attain causal and explanatory knowledge. A contrastive perspective requires theoretical models to provide justification for the choice of causal elements and argue why the chosen factors provide a better explanation (Pearl 2009). In order to evaluate the explanatory power of the HFE models,

we adopt Marchionni's (2012) three dimensions of explanatory power: contrastive force, explanatory breadth, and explanatory depth.

(1) Contrastive force entails justification of causal background, assumptions, and contrastive explanation of a phenomenon. False models have fairly limited contrastive force, in the sense that they handle some contrastive questions but not others (Morton 1990).
(2) Explanatory breadth indicates the extent to which a model accounts for different phenomena with the same or fewer explanatia. Explanatory breadth requires models' explanatia to be abstract enough to encompass a wider range of phenomena. Simply put, a model must be effectively generalizable to problems and populations beyond a single observation and occasion. Explanatory breadth is the matter of the unifying power of a model and whether a model can explain more of the phenomenon by encompassing different classes and instantiations of the phenomenon. The side effect of a high degree of explanatory breadth is the limited ability of the model to answer fine-grained questions about specific problems. On the flip side, models that aim to incorporate abundant information specific to a phenomenon in a particular occasion have limited unifying power but are better at answering fine-grained questions. Ultimately, selecting the right model depends on the interest and purpose of the study.
(3) Explanatory depth refers to the layers of investigation for underlying causal mechanisms. Achieving explanatory depth is typically a matter of describing mechanisms that component parts are at a lower level than the phenomenon to be explained (Hitchcock and Woodward 2003). However, the amount of information about the causal factors should not be confused with the depth of explanation. While deep explanations are often more detailed than shallow ones, detailed explanations are not always deep. A deep explanation discusses how the explanatory factors are responsible for the explanandum. Therefore, deep explanation requires theoretical and computational models to decompose their constructs and elaborate causal processes that give rise to specific behavior. Whether such elaboration takes place at a lower biological level or higher abstract level is mainly concerned with 'levels' problem, pertinent to the problem at hand and the level of analysis (Eronen 2021; Shapiro 2019).

Criterion 4: empirical adequacy

Empirical adequacy of a theory (or model or set of scientific claims) can be achieved when the claims about empirical phenomena are correct (Van Fraassen 1980; Bhakthavatsalam and Cartwright 2017). This requires the theory's assertions to be consistent with empirical evidence (Fawcett 2005). If the empirical findings corroborate the theoretical statements, it may be fair to tentatively accept the assertions as reasonable. If the empirical findings contradict the assumptions, it is reasonable to conclude that the assertions are incorrect. Empirical adequacy is different from the criterion of empirical testability as Empirical adequacy concerns the verisimilitude of a theorý s predictions, while empirical testability only refers to the extent to which a theory can be tested.

The propensity for circular reasoning should be noted while evaluating the model's empirical adequacy. If evidence is always evaluated in the context of a single model, it may be difficult to notice results that contradict that model. Indeed, if researchers repeatedly

expose, explain, and interpret data *via* the lens of a single model, the end result may be limited to the expansion of that model and that model alone (Ray 1990). Circular reasoning can be avoided by carefully examining the empirical findings to evaluate their degree of congruence with the model's ideas and propositions, as well as from the standpoint of competing models (Platt 1964). In other words, when interpreting evidence acquired considering a model, it is always necessary to take alternative models into consideration.

A single test of a model is unlikely to offer the conclusive evidence required to verify its empirical validity. As a result, all connected studies' conclusions should be considered when making decisions about empirical adequacy. To integrate the results of related investigations, meta-analysis, and other formal approaches can be employed. The goal of evaluating empirical adequacy is not to determine the absolute truth of the model but rather to identify the level of confidence received by the empirical evidence. The consequence of evaluating empirical adequacy is then a decision about whether one or more of the model's concepts or propositions need to be modified, refined, or discarded (Fawcett 2005). More importantly, since studies with incongruent results have more weight than studies with compatible results, empirical adequacy may also indicate how well a model manages disconfirming evidence. A model should provide an explanation for any discomforting instances (Gould 1991; Van de Ven 2007). It is also equally important to point out that it is not sufficient that only some parts of a model are congruent with empirical data rather, the entirety of a model must be empirically adequate and valid.

To evaluate the empirical adequacy criterion, a comprehensive review of the empirical research guided by the model must be performed. In this regard, the criterion can be stated as the questions:

(1) Are theoretical assertions made by the model congruent with empirical evidence?
(2) Has the entire model been tested in different studies?

Criterion 5: pragmatic adequacy/applicability

An applied field such as HFE is particularly concerned with practice and identifying theories that are most useful. HFE strives to improve the efficacy and efficiency of work and other activities, as well as human standards, including enhanced safety, reduced fatigue and stress, and improved quality of life (Sanders and McCormick 1998). Many scholars have claimed that knowledge transfer and synergy between HFE research and practice are required to attain these goals (Caple 2008; Meister 2018; Salas 2008; Sind-Prunier 1996). Getty (1995) emphasized the importance of HFE principles being based on robust and validated research, as well as the fact that the appropriate science and practice of HFE have long-term consequences for the discipline's future. Karwowski (2005) expanded on the significance of theory in the HFE field by identifying three primary paradigms: (1) HFE theory, which involves the ability to recognize, explain, and appraise human-system interactions; (2) HFE abstraction, which deals with those interactions to make predictions about the real world; and (3) HFE design, which involves utilizing the understanding about those interactions in order to design systems that can fulfil consumer needs and other necessary requirements.

Pragmatic adequacy is the extent to which a theory/model can offer effective solutions to real-world problems, which is based on the idea that theories are created to 'solve human

and technical problems and to improve practice' (Kerlinger 1979, 280). The pragmatic adequacy criterion requires that the application of a model is generally feasible to implement. Thus, it is expected that a good HFE model:

(1) recognizes the domain(s) to which it can be applied to,
(2) provides recommendations on how to implement the proposed model in that domain, and
(3) clarifies specific areas in which the model can provide useful and tangible results.

Criterion 6: recognizing humans as active agents

Witkin and Gottschalk (1988) argue that traditional theory evaluation criteria are not necessarily adequate to assess theories in social sciences and social work. They suggest theories should account for human beings as active agents. That is, humans are capable of reflecting on their own actions, overcoming distractions, making decisions, and adopting new principles and beliefs (Harré 1984). Thus, assumptions of people as mechanically responding to stimuli are less favourable than recognizing people as agents with their beliefs and intentions. People act, not simply behave. Such actions may impact the environment, change the course of events, and create new problem spaces (Øvergård, BjØrkli, and Hoff 2008). Viewing humans as active agents also shifts focus from exclusively identifying 'causes' of behaviour to the consequences of actions in a sociotechnical system. In this line, Gauch (2012) differentiates 'inference' and 'decision' problems. Despite a tight relationship, inference problems follow true beliefs, while decision problems follow ideal actions. Decision theory divides the causes of a situation into two distinct groups based on whether we have the power to control the cause or not (Gauch 2012). What we can control is the action or choice, and what we cannot control is the 'state.' Each combination of action and state provides an 'outcome' that has a specific utility or consequence that determines the value or benefit of the outcome. Since an uncontrollable situation (i.e. state) is usually unknown and changing, decision problems require Bayes inference to assess the probability of the state (based on prior and likelihood). Also, the response to the expected utility is not always linear. Decisions may have several criteria to be optimized simultaneously, possibly with some trade-offs and compromises. Therefore, inference and decision problems may have completely different solutions and outcomes.

Hence, a good HFE model recognizes humans as active agents and pursues modelling approaches that strive to explain the processes that give rise to human decisions, actions, and the meanings of future events (Kennedy 2012). So, the criterion for a good HFE model is:

(1) Does the model take human judgments, motivations, emotions, and socially driven behaviours into consideration?

Criterion 7: models of dynamic phenomena should be dynamic

Many problems in HFE cannot be reduced to a single static underlying cause but rather are emergent products of internal interactions in a complex socio-technical system (Guastello

2017). Complex systems constantly experience change as relationships and interconnections evolve and adapt to their dynamic environment (Dekker, Cilliers, and Hofmeyr 2011). Temporal patterns are the footprint of the dynamic environment. Time is also a fundamental element in modelling human-machine interaction (De Keyser, Decortis, and Van Daele 1988; Hollnagel 2002). This is because in dealing with systems and automation, humans must evaluate events in the limited time available, plan actions and execute them. Information required for this process also needs to be updated and checked regularly. Therefore, not only do mental processes and actions take time, but different time frames also demand the prioritization of concurrent activities (Hollnagel 2002). It is of interest to understand whether an HFE model can address the dynamic behaviour of a phenomenon or not. To evaluate this criterion, we seek to uncover whether the model explicitly indicates time as an essential component of a dynamic construct or not. The indicator of this criteria is as follows:

(1) If the phenomenon is dynamic, does the model acknowledge time as a variable?

Thus far, we have proposed seven different criteria with a number of indicators. Table 1 provides a summary of the proposed criteria for model evaluation as well as the indicators for each criterion.

In the following sections, we examine some of the prominent models of Trust in Automation (TiA) according to the proposed criteria.

Assessing models of trust in automation

Trust is an abstract, complex, and multidimensional concept that can be attributed to wide-ranging entities such as humans, machines, organizations, institutions, and countries (Abbass et al. 2016). In the context of human-automation interaction (HAI), trust is acknowledged to be an essential element in the use, misuse, or disuse of automation (Parasuraman and Riley 1997). Trust is not all or nothing but is a continuous phenomenon that can be attributed to an agent as a whole or to specific parts, capabilities, or functions of that agent (Hou, Ho, and Dunwoody 2021; Chiou and Lee 2023). Also, trust is situation and task-dependent, which means it can vary even towards the same agent at different occasions and times. For instance, one may fully trust his/her partner, but not in specific tasks like cooking. Trust has been treated as both a relatively static and dynamic phenomenon. As a psychological construct, trust has a long-term propensity that is relatively stable until it is broken (Jarvenpaa, Knoll, and Leidner 1998; Mayer, Davis, and Schoorman 1995), but it can also change, evolve, and degrade over time (Desai et al. 2013; Schaefer 2013; Wilson, Straus, and McEvily 2006). Research also points out asymmetry between development and loss of trust over time, meaning that the process of building trust is slow and steady while distrust can happen quickly by a single event or inconsistency in trustee's behaviour (Burt and Knez 1996; Lewicki and Bunker 1996; Gambetta 1988). This asymmetry has made some scholars treat trust and distrust as two distinct constructs that can evolve or decline independently (Kramer, Brewer, and Hanna 1996; Lewicki, McAllister, and Bies 1998).

More than three decades of human-automation interaction research have resulted in the emergence of numerous theories and models, endeavouring to provide insight into human performance within complex sociotechnical systems. Modelling trust in automation has

Table 1. Criteria for model evaluation in HFE.

Criteria	Indicator(s)	Reference
(C1) Testability/Falsifiability	(1) Can the model be operationalized? Is there a way of measuring the components and constructs in the theory? (2) Does the model/theory propose research design for testing the model's assumptions? (3) Are the tools and data analysis techniques adequate to measure the model propositions?	Popper (1969), Cramer (2013), Fawcett (1988), Silva (1986)
(C2) Predictive power	Can the model make predictions about: (1) Existence of effect? (2) Direction (or sign) of effect? (3) Direction and interval estimate of effect? (4) Mathematical specification of predicted effect?	Meehl (1967), Dienes (2008), Meehl (1978), Velicer et al. (2008), Freedman (2010), McElreath (2018)
(C3) Explanatory power	Does the model provide (1) Contrastive force? (2) Explanatory breadth? (3) Explanatory depth?	Cramer (2013), Prochaska, Wright, and Velicer (2008), Garfinkel (1982), Lipton (1990), Ylikoski (2007). Marchionni's (2012), Morton (1990), Hitchcock and Woodward (2003)
(C4) Empirical adequacy	Are theoretical assertions made by the model congruent with empirical evidence? Has the entire model been tested in different studies?	Van Fraassen (1980), Bhakthavatsalam and Cartwright (2017), Fawcett (2005), Gould (1991), Van de Ven (2007)
(C5) Pragmatic adequacy	Does the model: (1) recognize the domain(s) to which it can be applied to? (2) provides recommendations on how to implement the proposed model in that domain? (3) clarify specific areas in which the model can provide useful and tangible results?	Getty (1995), Karwowski (2005), Caple (2008), Meister (2018), Salas (2008), Sind-Prunier (1996)
(C6) Human as active agent	Does the model take human judgments, motivations, emotions, and socially driven behaviours into consideration?	Witkin and Gottschalk (1988), Gauch (2012), Kennedy (2012)
(C7) Dynamic properties	If the phenomenon is dynamic, does the model acknowledge time as a variable?	Guastello (2017), Dekker, Cilliers, and Hofmeyr (2011), De Keyser, Decortis, and Van Daele (1988), Hollnagel (2002)

undergone various modelling attempts ranging from regression models, time-series models, qualitative models, argument-based probabilistic models, and neural net models with each modelling approach having its pros and cons (Moray and Inagaki 1999). Regression-based models are useful in identifying the independent and dependent variables, as well as the relationships among them, hence providing rigid testability and predictive power. These models, however, are unable to capture the dynamic variances in trust formation and can only be used for factors that influence trust which do not significantly vary during interaction with automation. Time-series models are used to capture the dynamic relationship between trust and other independent variables, but they require prior knowledge about the causal factors and large enough data for validation (Moray and Inagaki 1999; Desai 2012). Argument-based probabilistic trust models are based on information value theory and utilize evidence to lower the degree of uncertainty in the model's outputs. The output of

the model is the probability that a particular course of action will succeed, i.e. how much one can trust the decision aids suggestions (Cohen et al. 1997). Neural net models are data-driven models. They can make accurate predictions about trust and control allocation strategies but due to the nature of such models (varying coefficients from one data set to another), it is not feasible to extract a meaningful explanation about how the model works. Neural nets are not models of psychological processes but rather predictive models applied in human-machine systems (Moray and Inagaki 1999).

Data collection

To identify the existing models of trust in automation, four databases were searched: Web of Science, Scopus, ScienceDirect, and Google Scholar. This led to several duplications but also ensured thorough indexing of academic databases. The search was restricted to the title, abstract, and keywords of the publications using the search string: ('Trust in Automation' OR 'Trust in Automated' OR 'Trust in Autonomy' OR 'Trust in Autonomous' OR 'Trust in Robots') AND ('Model*'). Additionally, we examined the literature review articles on trust in automation models (e.g. French, Duenser, and Heathcote 2018; Abbass, Scholz, and Reid 2018; Adams, Bruyn, and Houde 2003; Hussein, Elsawah, and Abbass 2020) and employed snowball approach to ensure inclusion of all relevant studies. The initial screening was performed to remove any duplicates. The second-stage screening of articles required analysing the abstracts to identify whether the study potentially proposes a model of trust in automation. At the second-stage screening, we made some scoping constraints to exclude works focused on just one component (e.g. the effect of culture on TiA) and/or studies that only peripherally mentioned trust in automation.

After a comprehensive review of the articles, thirty-six studies were selected for evaluation. The studies are classified into two main clusters. The first cluster of models involves theoretical research intending to offer conceptual models of trust in automation which share many similarities. They often provide causal factors related to the automation, the individual, and to the environment's characteristics and are generally presented in a network diagram. Conceptual models consider trust as a mediator of the operator's reliance on automation. The second cluster of studies involves computational models, aimed at providing mathematical and/or probabilistic models that can predict trust by incorporating causal factors and relationships among them.

Criteria weighting

To evaluate the models of trust in automation, it is important to arrange the proposed criteria according to a ranking system. This is because different criteria have relative importance in model evaluation. A model can be portrayed as dynamic and suggest a pragmatic application, and yet unfalsifiable. Conversely, a testable model can lack temporal property and/or have limited predictive/explanatory power. Therefore, identifying the relative weight of each criterion seems necessary. The model evaluation can be seen as a Multi-criteria decision-making (MCDM) problem. For this purpose, this study utilized the Best Worst Method (BWM) as a branch of MCDM. The BWM uses ratios of the relative importance of criteria in pairwise comparisons specified by the decision-maker (Liang, Brunelli, and Rezaei 2020). Compared to other MCDM methods, such as Analytical Hierarchy Process

(AHP), BWM requires fewer comparison data for generating consistent pairwise comparisons (Rezaei 2015, 2016). The BWM starts with identifying the most and least important criteria, followed by ratings for the relative importance of other criteria in pairwise comparisons with the most and least important ones. To derive the weights of each criterion, two independent researchers followed the standard steps in BWM, as described below. The overall weighting is then calculated as the mean from the two evaluations.

Step 1 is to determine a set of decision criteria as $\{C_1, C_2, \ldots, C_n\}$. The decision criteria in this study can be shown as:

$$\{Testability\,(C_1), Predictive\,Power\,(C_2), \ldots, Dynamic\,Properties\,(C_7)\}$$

Step 2 is to define the most and least important criteria. In this study, testability and pragmatic adequacy are considered the most and least important criteria, respectively. This is because if a model is not testable, there is no practical way to examine many of the remaining criteria. However, a model can pass some essential criteria and is yet to be applied in real-world settings.

Step 3 is to decide the importance of the best criterion over all other criteria using a scale from 1 to 9. The result would be a vector as:

$$A_B = (\alpha_{B1}, \alpha_{B2}, \ldots, \alpha_{Bn})$$

Where α_{Bj} denotes the importance of the best criterion B over criterion j.

Step 4 is to decide the importance of all the criteria over the worst criterion using a scale from 1 to 9. The result would be a vector as:

$$A_w = (\alpha_{1w}, \alpha_{2w}, \ldots, \alpha_{nw})^T$$

Where α_{jw} denotes the importance of the criterion j over the worst criterion W.

Step 5 is to determine the optimal weights vector $(W_1^*, W_2^*, \ldots, W_n^*)$, where for each pair of W_B/W_j and W_j/W_w, there is $W_B/W_j = \alpha_{Bj}$ and $W_j/W_w = \alpha_{jw}$. To satisfy these conditions for all j, the below linear min-max problem must be solved according to the following formula:

$$\min \max \left\{ \left| \frac{W_B}{W_j} - \alpha_{Bj} \right|, \left| \frac{W_j}{W_w} - \alpha_{jw} \right| \right\}$$

Subject to

$$\sum_j W_j = 1$$

Using the BWM Excel solver (Rezaei 2022), the relative weight of each criterion is calculated as shown in Table 2.

Table 2. BWM criteria weighting.

Criteria Number = 7	Criterion 1	Criterion 2	Criterion 3	Criterion 4	Criterion 5	Criterion 6	Criterion 7
Names of Criteria	Testability	Predictive power	Explanatory power	Empirical adequacy	Pragmatic adequacy	Human as Active Agent	Dynamic Properties
Select the Best	Testability						
Select the Worst	Pragmatic adequacy						
Best to Others	Testability	Predictive power	Explanatory power	Empirical adequacy	Pragmatic adequacy	Human as Active Agent	Dynamic Properties
Testability	1	1	5	8	9	9	9
Others to the Worst	Pragmatic adequacy						
Testability	9						
Predictive power	6						
Explanatory power	7						
Empirical adequacy	3						
Pragmatic adequacy	1						
Human as Active Agent	2						
Dynamic Properties	3						
Weights	Testability	Predictive power	Explanatory power	Empirical adequacy	Pragmatic adequacy	Human as Active Agent	Dynamic Properties
	0.392190465	0.299651141	0.10135259	0.06334537	0.03084644	0.056306996	0.056307

Model evaluation

After identifying the weight of each criterion, the evaluation is carried out for the degree to which a model can satisfy each criterion. The models are rated on a subjective scale from 1 to 9 for each criterion, normalized ($X_{norm(i,j)}$), and computed the overall scores (OS_i) as:

$$X_{norm(i,j)} = \frac{X_{(i,j)}}{\max X_j}$$

$$OS_i = \sum (X_{norm(i,j)} * W_j)$$

Where $X_{(i,j)}$ is a degree to which model i can satisfy the criterion j, X_j is the j^{th} column of matrix X, and W_j is the relative weight of criterion j.

Furthermore, a second assessment is conducted for a random 20% of the models (four conceptual and three computational) to realize the reliability of the evaluation. Subsequently, the inter-rater reliability as a measure of agreement among evaluations (Krippendorff 2011, 2004) is calculated with Krippendorff's $\alpha_k = 0.88$ which signifies an acceptable inter-rater score.

To demonstrate the evaluation process, Muir's (1987) conceptual model of trust is selected as an illustrative example. The model draws upon trust taxonomies proposed by Barber (1983) and Rempel, Holmes, and Zanna (1985), and encompasses the expectation of persistence, technically competent performance, and fiduciary responsibility. Since the model does not specify the ways to operationalize and measure its components, the testability of the entire model becomes restricted. However, the linear regression-based formulation indicates a resealable predictive ability of the model. The model receives a low explanatory power score as it fails to provide sufficient explanatory depth/breadth despite its attempts to distinguish itself (i.e. contrastive force) from the previous interpersonal trust models. The empirical adequacy of the model is also fairly limited to the experimental studies of trust and human intervention in a process control simulation (Muir and Moray 1996). With regard to the pragmatic adequacy criterion, the model provides some generic recommendations about the calibration of trust for decision support systems. However, it falls short in specifying the applicable domains and the practical benefits of using the model. Additionally, the model also does not adequately account for humans' judgments, biases, and socially driven behaviours resulting in a low score in this area. Although Muir's (1987) model discusses trust as a dynamic phenomenon, it cannot be considered as a dynamic model since it fails to explain the temporal characteristics of trust in automation.

Results

The evaluation of TiA models was conducted based on the proposed criteria to assess their adherence to each criterion. Prior to discussing the evaluation results, it is essential to examine the relationships between the criteria. As illustrated in Table 3, there exists a positive correlation between the testability and predictive power of the models. This is because in order to measure the predictive power, the model's assumptions must be measurable and testable. Testability is also a meaningless idea without the model generating some predictions to be tested. Conversely, explanatory power and predictive power appear to be inversely

Table 3. Correlational values among seven criteria.

		C1-Testability	C2-Predictive Power	C3-Explanatory Power	C4-Empirical Adequacy	C5-Pragmatic Adequacy	C6-Human Agency	C7-Dynamic Properties
Conceptual Models	C1	1.00						
	C2	**0.66**	1.00					
	C3	−0.04	−0.01	1.00				
	C4	0.28	0.28	0.21	1.00			
	C5	0.59	0.33	0.06	0.14	1.00		
	C6	−0.30	−0.37	0.61	0.05	0.18	1.00	
	C7	0.58	0.45	0.30	0.08	0.53	0.18	1.00
Computational Models	C1	1.00						
	C2	**0.81**	1.00					
	C3	−0.44	−0.39	1.00				
	C4	0.25	0.22	0.22	1.00			
	C5	0.52	0.64	−0.34	0.37	1.00		
	C6	0.51	0.53	−0.27	0.40	0.44	1.00	
	C7	0.68	0.63	−0.23	0.28	0.42	0.60	1.00

correlated. This can be understood from a perspective of modelling functionality and the trade-off between the explanation and prediction (Watts et al. 2018; Hofman, Sharma, and Watts 2017; Yarkoni and Westfall 2017). Conceptual causal models that aim to encompass a wide range of instances by incorporating ample causal factors may have limited predictive capabilities. On the other hand, predictive models (e.g. regression, time-series) may achieve higher accuracy by narrowing down the causal elements, resulting in less generalizable outcomes (i.e. reduced explanatory power).

Criterion 1, testability

With regards to the testability criterion, the components of early conceptual models are often expressed in generic terms such as ability, benevolence, integrity (Mayer, Davis, and Schoorman 1995), faith, and personal attachments (Madsen and Gregor 2000). The generic terminology reduces the possibility of the models being operationalized and tested and therefore defies the testability criterion. A number of studies provide mathematical notations (Muir 1994) regression-based (Muir 1994; Lee and Moray 1992), and time series (Lee and Moray 1994), but these can be seen as partial representations of the original conceptual models. Computational models, on the other hand, offer more precise and quantifiable definitions for models' variables in order to be validated with data, and hence perform better in this criterion.

Criterion 2, predictive power

With respect to predictive power, most conceptual models can provide the existence of effect (sub-criterion C2-1). Muir (1994) offers a linear regression formulation as a mathematical specification of predicted effect (sub-criterion C2-4). Similarly, Lee and Moray (1992) Autoregressive Moving Average Vector (ARMAV) model receives a higher score in the predictive power criterion. The computational models that are expressed using mathematical equations have normally a higher predictive ability. However, Sheridan's (2019) three models of signal detection, statistical parameter estimation, and model-based control as well as the system dynamics model proposed by Hussein, Elsawah, and Abbass (2019) do not offer sufficient details for the variables and therefore generate less risky predictions.

Criterion 3, explanatory power

Explanatory power is evaluated for the degree to which a model can provide contrastive force, explanatory breadth, and explanatory depth. To do so, the theoretical assumptions of the models were reviewed to identify whether the model justifies the choices for its components/parameters, the relationships between the components, and the relative advantage of the model compared to previous models. Moreover, we sought to consider whether the model attempted to decompose and elaborate its structural elements and answer 'how' questions (explanatory depth). The model's assumptions are also examined for conceivable generalizability (explanatory breadth).

A higher level of abstraction in conceptual models allows for encompassing a wider range of phenomena. Models of Lee and See (2004), Hoff and Bashir (2015), and Hancock et al. (2011) received the highest scores in this criterion for providing an ample contrastive force and justification of assumptions while offering a broad explanatory breadth to

encompass a wider range of TiA instances. However, these models (and many other conceptual models) have a relatively shallow explanatory depth in decomposing the underlying causal mechanisms and explaining the interactions that give rise to TiA. Among computational models, the extended decision field theory model (Gao and Lee 2006) provides a detailed explanation and highlights the inertia of trust, the nonlinear relationship between trust, self-confidence, and reliance on automation in a closed-loop dynamic model.

Criterion 4, empirical adequacy

The empirical adequacy of the models is examined to realize whether the model's assertions are supported by empirical research. Several studies have acknowledged the role of different factors on TiA, such as age (Ho et al. 2005), personality traits (Merritt and Ilgen 2008; Szalma and Taylor 2011), culture (Huerta, Glandon, and Petrides 2012), gender (Nomura et al. 2008), self-confidence (de Vries, Midden, and Bouwhuis 2003), and automation reliability (Parasuraman and Riley 1997; Dzindolet et al. 2003). Nonetheless, the empirical adequacy of the conceptual models remain somewhat limited. In our assessment, the meta-analysis model proposed by Hancock et al. (2011) receives a higher score for offering an evidence-based model of TiA, although the entirety of the model has yet to undergo comprehensive testing. Similarly, the empirical adequacy of the computational models is typically constrained to data fitting and model validation within a single study.

Criterion 5, pragmatic adequacy

Pragmatic adequacy pertains to the application of TiA models in real-world settings. This criterion requires the TiA models to explicitly specify the domain(s) to which they are applicable. Models that are specifically tailored to a particular context excel in this criterion, as they are primarily designed for a specific setting. For instance, Kraus et al. (2020) model is mainly developed for automated driving (AD) vehicle systems and offers new insights into the processes involved in trust calibration prior to and during the take-over request (TOR). Argument-based Probabilistic Trust (APT) model (Cohen et al. 1997) explores its feasibility to be implemented in a military decision-aiding environment for Rotorcraft Pilot's Associate (RPA). Among computational models, those that aimed to be utilized in real-world applications such as human-robot interactions (e.g. Xu and Dudek 2015, 2012), or automated driving systems (Azevedo-Sa et al. 2021) receive higher scores in terms of pragmatic adequacy.

Criterion 6, humans as active agents

Humans are self-reflecting actors that do not mechanically respond to stimuli but rather reflect, draw on previous experience, make choices, and anticipate the outcome of their decisions. The 'Humans as active agents' criterion requires the TiA models to take human judgment, biases, motivations, emotions, and socially driven behaviour into consideration. For example, Cohen et al. (1997) model incorporates different levels of operators' understanding of automation trustworthiness by integrating an event tree model that represents various pathways denoting different scenarios in which an operator may need decision support.

Among computational models, Hoogendoorn et al. (2013) introduced an adaptive biased-based trust model that is designed to perform in situations where humans have to make

decisions to trust one of the multiple heterogeneous trustees. The model considers human inclinations to an agent system based on available cues and previous interactions with the system. In another study, Akash et al. (2017) proposed a third-order linear trust model that can capture the cumulative perception of trust as well as bias in human's expectation of a particular interaction with automation.

Criterion 7, dynamic criterion

Walker, Stanton, and Salmon (2016, 5) describe trust as 'a dynamic phenomenon, moving along a continuum,…'. The dynamic criterion stipulates that if a phenomenon is dynamic, the models representing it should also be dynamic and capable of explaining the phenomenon in a dynamic manner. While computational models have the advantage of producing time-series and simulation models, conceptual models can provide a dynamic understanding of evolution and degradation of trust by elucidating how time as a variable plays a role in the modelling process. In our evaluation, we assessed the extent to which existing models consider time as a parameter. This process takes a range of forms; from the inclusion of information feedback loops, describing temporal dynamics of trust, to the development of time-series and dynamic simulation models.

Lee and Moray (1992) time-series model represents an early attempt to highlight the temporal characteristics of trust. The dynamic model accounts for a greater amount of variance compared to a simple regression model (79.1% versus 53.3%), also indicating its improved predictive power. Lee and See (2004) and Hoff and Bashir (2015) models are also notable in reflecting the dynamics of trust through signifying closed feedback loops and the distinction between initial and dynamic learned trust during human-automation interaction. Building upon the assumptions of these two models, Kraus et al. (2020) proposed a theoretical model to capture the dynamics of trust calibration in highly automated driving settings. Another contribution is the introduction of a real-time computational model of trust for human-automation collaboration called trust-POMDP, which integrates measured trust in the automation decision-making (Chen et al. 2018). In a different approach, Gao and Lee (2006) proposed a model based on the extended decision field theory (EDFT) to capture the dynamics and nonlinear characteristics of trust.

Tables 4 and 5 summarize the results for theoretical and computational models in all the criteria.

Discussion

The model evaluation revealed key differences between TiA models. Three conceptual models particularly stood out in terms of their overall scores. Lee and See (2004) model is remarkable in providing a widely accepted definition of trust in automation and a closed-loop dynamic framework that governs trust and its impact on reliance. The model considers various causal factors underlying trust in automation including information assimilation and belief formation, individual, organizational, cultural, and environmental context. Despite the limitation in operationalization and testability of the model's assumptions, Lee and See (2004) model is notable in elucidating the dynamic evolution of trust and the dimensions that describe the basis of trust. Desai's (2012) qualitative model of trust in autonomous robot teleoperation represents an important step in using the Area Under Trust

Table 4. Summary Scores of TiA models (conceptual and computational).

Model/Criteria	C1	C2	C3	C4	C5	C6	C7	Overall Score
BWM pairwise weight	0.392	0.300	0.101	0.063	0.031	0.056	0.056	
Muir (1987)	5	6	3	3	1	1	1	4.40
Lee and Moray (1992)	6	6	3	2	3	2	6	5.12
Muir (1994)	6	4	3	2	1	1	2	4.18
Cohen et al. (1997)	6	4	3	2	4	3	2	4.39
Madsen and Gregor (2000)	3	3	4	1	2	3	1	2.83
Seong and Bisantz (2000)	5	4	3	2	2	1	1	3.76
Kelly et al. (2001)	5	6	3	2	2	2	1	4.42
Adams, Bruyn, and Houde(2003)	5	5	6	2	1	3	3	4.56
Nickerson and Reilly (2004)	4	5	3	1	2	1	2	3.67
Lee and See (2004)	**6**	**6**	**8**	**3**	**3**	**3**	**4**	**5.64**
Madhavan and Wiegmann (2004)	4	3	4	3	2	3	2	3.41
Hancock et al. (2011)	4	4	8	4	2	3	2	4.17
Desai (2012)	6	6	7	2	3	2	3	5.36
Chien et al. (2014)	4	4	7	2	2	3	1	3.89
Hoff and Bashir (2015)	5	5	8	2	2	3	4	4.85
Bindewald, Rusnock, and Miller (2018)	3	4	4	2	1	3	1	3.16
Kraus et al. (2020)	6	5	6	2	4	3	5	5.16
Hou, Ho, and Dunwoody (2021)	4	3	7	1	1	3	2	3.55
Solberg et al. (2022)	3	3	6	1	1	3	2	3.06
Gao and Lee (2006)	8	8	7	2	5	5	7	7.20
Itoh (2011)	5	6	7	2	3	2	2	4.91
Xu and Dudek (2012)	7	7	6	2	6	6	8	6.55
Gao et al. (2013)	8	8	7	2	3	3	8	7.08
Hoogendoorn et al. (2013)	8	8	5	2	2	8	7	7.07
Xu and Dudek (2015)	8	8	4	2	6	3	5	6.70
Sadrfaridpour et al. (2016)	8	8	6	2	4	3	7	6.96
Akash et al. (2017)	8	8	6	3	6	7	8	7.36
Hu et al. (2019)	**8**	**8**	**7**	**3**	**6**	**7**	**8**	**7.46**
Akash, Reid, and Jain (2018)	8	8	5	2	6	7	8	7.20
Chen et al. (2018)	8	6	5	2	3	4	7	6.28
Hussein, Elsawah, and Abbass (2019)	5	5	7	2	2	3	6	4.86
Sheridan (2019)	5	5	7	2	2	3	3	4.69
Nam et al. (2020)	8	8	6	2	6	7	6	7.19
Guo and Yang (2021)	7	8	5	2	5	7	8	6.77
Chen et al. (2020)	7	8	6	2	4	5	8	6.73
Azevedo-Sa et al. (2021)	7	8	5	2	6	4	7	6.58

Table 5. Normalized summary scores of TiA models (conceptual and computational).

Model/Criteria	C1	C2	C3	C4	C5	C6	C7	Overall Score
BWM pairwise weight	0.392	0.300	0.101	0.063	0.031	0.056	0.056	
Muir (1987)	0.63	0.75	0.38	0.75	0.17	0.13	0.13	0.57
Lee and Moray (1992)	0.75	0.75	0.38	0.50	0.50	0.25	0.75	0.66
Muir (1994)	0.75	0.50	0.38	0.50	0.17	0.13	0.25	0.54
Cohen et al. (1997)	0.75	0.50	0.38	0.50	0.67	0.38	0.25	0.57
Madsen and Gregor (2000)	0.38	0.38	0.50	0.25	0.33	0.38	0.13	0.36
Seong and Bisantz (2000)	0.63	0.50	0.38	0.50	0.33	0.13	0.13	0.49
Kelly et al. (2001)	0.63	0.75	0.38	0.50	0.33	0.25	0.13	0.57
Adams, Bruyn, and Houde(2003)	0.63	0.63	0.75	0.50	0.17	0.38	0.38	0.59
Nickerson and Reilly (2004)	0.50	0.63	0.38	0.25	0.33	0.13	0.25	0.47
Lee and See (2004)	**0.75**	**0.75**	**1.00**	**0.75**	**0.50**	**0.38**	**0.50**	**0.73**
Madhavan and Weigmann (2004)	0.50	0.38	0.50	0.75	0.33	0.38	0.25	0.45
Hancock et al. (2011)	0.50	0.50	1.00	1.00	0.33	0.38	0.25	0.56
Desai (2012)	0.75	0.75	0.88	0.50	0.50	0.25	0.38	0.69
Chien et al. (2014)	0.50	0.50	0.88	0.50	0.33	0.38	0.13	0.50
Hoff and Bashir (2015)	0.63	0.63	1.00	0.50	0.33	0.38	0.50	0.62
Bindewald, Rusnock, and Miller (2018)	0.38	0.50	0.50	0.50	0.17	0.38	0.13	0.41
Kraus et al. (2020)	0.75	0.63	0.75	0.50	0.67	0.38	0.63	0.67
Hou, Ho, and Dunwoody (2021)	0.50	0.38	0.88	0.25	0.17	0.38	0.25	0.45
Solberg et al. (2022)	0.38	0.38	0.75	0.25	0.17	0.38	0.25	0.39
Gao and Lee (2006)	1.00	1.00	0.88	0.50	0.83	0.63	0.88	0.92
Itoh (2011)	0.63	0.75	0.88	0.50	0.50	0.25	0.25	0.63
Xu and Dudek (2012)	0.88	0.88	0.75	0.50	1.00	0.75	1.00	0.84
Gao et al. (2013)	1.00	1.00	0.88	0.50	0.50	0.38	1.00	0.91
Hoogendoorn et al. (2013)	1.00	1.00	0.63	0.50	0.33	1.00	0.88	0.90
Xu and Dudek (2015)	1.00	1.00	0.50	0.50	1.00	0.38	0.63	0.86
Sadrfaridpour et al. (2016)	1.00	1.00	0.75	0.50	0.67	0.38	0.88	0.89
Akash et al. (2017)	1.00	1.00	0.75	0.75	1.00	0.88	1.00	0.95
Hu et al. (2019)	**1.00**	**1.00**	**0.88**	**0.75**	**1.00**	**0.88**	**1.00**	**0.96**
Akash, Reid, and Jain (2018)	1.00	1.00	0.63	0.50	1.00	0.88	1.00	0.92
Chen et al. (2018)	1.00	0.75	0.63	0.50	0.50	0.50	0.88	0.80
Hussein, Elsawah, and Abbass (2019)	0.63	0.63	0.88	0.50	0.33	0.38	0.75	0.63
Sheridan (2019)	0.63	0.63	0.88	0.50	0.33	0.38	0.38	0.61
Nam et al. (2020)	1.00	1.00	0.75	0.50	1.00	0.88	0.75	0.92
Guo and Yang (2021)	0.88	1.00	0.63	0.50	0.83	0.88	1.00	0.87
Chen et al. (2020)	0.88	1.00	0.75	0.50	0.67	0.63	1.00	0.86
Azevedo-Sa et al. (2021)	0.88	1.00	0.63	0.50	1.00	0.50	0.88	0.85

Curve (AUTC) measure to account for an individual's long-term interaction experience with the robot. While the model was developed based on experimental data, it is not suitable for accurately predicting trust and human performance. Kraus's (2020) three-stage trust framework integrates the key assumptions of Lee and See (2004) and Hoff and Bashir (2015) trust models, providing a more detailed specification of the psychological processes involved in the formation and calibration of trust. The model distinguishes between the factors influencing trust prior to and during interactions, enabling a clearer understanding of interactions among various individual and situational processes. However, the model appears to overlook human agency and trusting behaviour for reliance on automation. Regarding computational models, Gao and Lee (2006) model of extended decision field theory (EDFT) and dynamic model of human-machine trust (Hu et al. 2019) are noteworthy for providing a testable, predictive, and dynamic explanation of trust in automation. These models excel in identifying the significance of cumulative trust and expectation bias.

Assuming a model could perfectly fulfil all the proposed criteria would be irrational as different models can vary in their performance across the seven criteria. A model may excel in one criterion while performing poorly in another. That is why some prefer the term 'ideals' rather than criteria for model evaluation (Van Lange 2013). That said, computational models tend to perform better in terms of the overall model scores. This is due to their testability and inclusion of articulated equations that allow for the inclusion of dynamic properties thereby enhancing their predictive power. Nonetheless, computational models are constrained by the causal factors included in the model which can limit their explanatory breadth and generalizability. As Hu et al. (2019) report, factors such as demographics, false alarms, misses, and the effect of past experience on the future trust level are often overlooked in the computational models.

A nonparametric statistical test reveals the key differences between the conceptual and computational models in fulfilling the criteria. As shown in Table 6, computational models generally outperform conceptual models in all criteria except criterion 3 (explanatory power) and criterion 4 (empirical adequacy). This is not surprising since conceptual models are typically designed to be more generalizable for a wide range of instances, thereby providing a broader explanatory scope. The qualitative nature of the conceptual models also allows for the inclusion of more causal factors, extensive explanation, and justification of model parameters, resulting in a higher contrastive force. The greater explanatory breadth and contrastive force in the conceptual models provide a general framework for empirical studies. Though not always the entirety of the model, certain assumptions have undergone empirical testing and validation. That being said, empirical adequacy received the lowest score among both conceptual and computational models, indicating a lack of empirical validation beyond a single study.

To summarize, while conceptual models offer valuable insight into how trust, reliance, and other factors may interact, their heuristic nature hinders accurate predictions regarding

Table 6. Nonparametric tests of TiA models.

Test Statistics[a]	C1	C2	C3	C4	C5	C6	C7	Overall Score
Mann-Whitney U	25.500	19.000	120.000	147.500	39.500	43.500	17.000	13.000
Wilcoxon W	215.500	209.000	310.000	337.500	229.500	233.500	207.000	203.000
Z	−4.400	−4.643	−1.341	−.564	−3.958	−3.964	−4.640	−4.706
Asymp. Sig. (2-tailed)	<.001	<.001	.180	.573	<.001	<.001	<.001	<.001
Exact Sig. [2*(1-tailed Sig.)]	<.001[b]	<.001[b]	.196[b]	.661[b]	<.001[b]	<.001[b]	<.001[b]	<.001[b]

[a]Grouping Variable: Model Type.
[b]Not corrected for ties.

trust and control allocation (Desai 2012). The use of general terminology in conceptual models poses a challenge for precise operationalization, limiting the testability and empirical validation of these models. This entails that there cannot be any observation that could possibly contradict the model's assumptions and refute them. Despite some consensus on the key factors influencing trust in automation, there remains no agreement on 'how' various factors and attributes combine into a single vector within existing TiA models (Sheridan 2019). This modelling challenge highlights the importance of the model's structure (Hollnagel 2002). Conceptual models tend to assume the interactions between various constructs and factors as unidirectional linear pathways. However, this stimulus-response logic, prevalent in both theories and experiments, greatly underestimates the complexity of the coupling effect between human agents, automation, and the environment (Kugler and Turvey 2015; Jagacinski and Flach 2018). Trust, as an outcome of prolonged interaction with automation on an infinite number of occasions, is far more complex to be modelled in a linear stimulus (cause) and response (effect) fashion. Failure in automation has a decaying reminiscence effect on future trust. On top of that, properties in dynamic systems can be induced by changes in other properties, resulting in simultaneous and reciprocal alterations (Van Gelder and Port 1995). This implies that changes in trust, which can be influenced by factors like automation reliability, may indirectly impact automation reliability itself through reliance on automation and intervening behaviours. The intrinsic complexity of sociotechnical systems introduces new complications that require a comprehensive consideration of the direction of causality and temporal priority of the causal variables (Jagacinski and Flach 2018; Guastello 2017; Van de Ven 2007). Therefore, efforts should be directed towards refreshing our epistemological understanding of complex systems and adopting novel modelling techniques that can accommodate the ever-growing complexity of socio-technical systems.

On a related note, and to address the question raised in the introduction section, a regression analysis was performed for the thirty-six models of trust in automation. By doing so, we aimed to gain insights into the temporal evolution of TiA research and assess the TiA progress over time. Figure 1 illustrates that the TiA models exhibit an upward trend, indicating a gradual advancement in the field. However, when considering the model's type as a covariate in the regression analysis (Table 7), it becomes evident that there is no

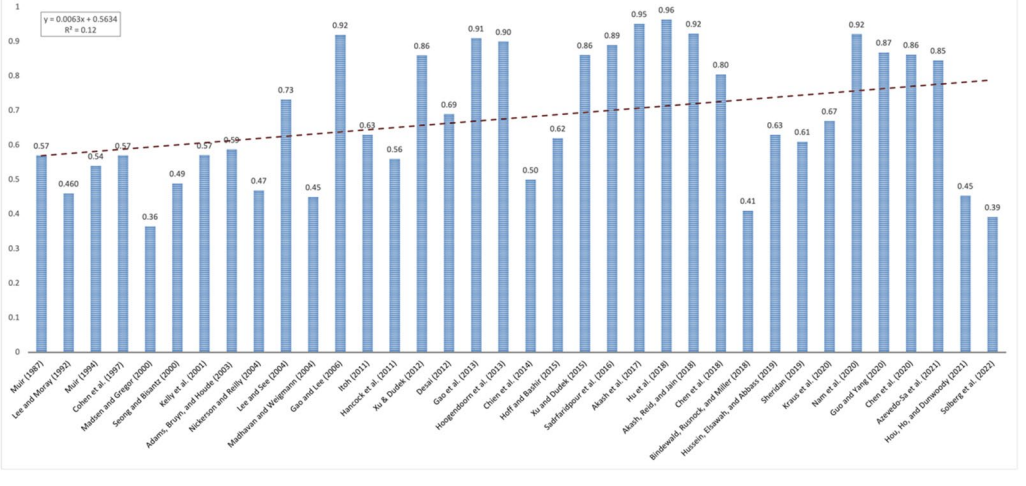


Figure 1. TiA model scores over time.

Table 7. Regression analysis of TiA models.

Regression Statistics*								
Multiple R	0.433782457		ANOVA					
R Square	0.18816722			*df*	*SS*	*MS*	*F*	*Significance F*
Adjusted R Square	0.164289785		Regression	1	0.239016301	0.239016301	7.88054589	0.008214854
Standard Error	0.174154864		Residual	34	1.03121717	0.030329917		
Observations	36		Total	35	1.27023347			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	−17.09374518	6.331220186	−2.699913237	0.010729251	−29.96033264	−4.227157724	−29.96033264	−4.227157724
Year	0.008838141	0.003148348	2.807231	**0.008214854**	0.002439927	0.015236355	0.002439927	0.015236355
*Regression Statistics***								
Multiple R	0.832128144		ANOVA					
R Square	0.692437247			*df*	*SS*	*MS*	*F*	*Significance F*
Adjusted R Square	0.673797081		Regression	2	0.879556968	0.439778484	37.14758854	3.55555E-09
Standard Error	0.108805707		Residual	33	0.390676502	0.011838682		
Observations	36		Total	35	1.27023347			
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	1.880208186	4.722284449	0.398156487	0.693079154	−7.727351764	11.48776814	−7.727351764	11.48776814
Model Type	−0.318147595	0.043252077	−7.355660535	1.90644E-08	−0.406144608	−0.230150583	−0.406144608	−0.230150583
Year	−0.000513705	0.002342093	−0.219335675	**0.827739614**	−0.005278729	0.00425132	−0.005278729	0.00425132

*Time as a covariate.
**Model type as a covariate.

significant change in TiA models over time. Thus, relying solely on a simple regression analysis with time as the covariate can be misleading. The observed reason for the upward trend can be attributed to the increased prevalence of computational models in recent years and not because of meaningful development in the TiA research programme.

Concluding remarks

For human factors and ergonomics (HFE) to progress as a scientific discipline, it is necessary to produce and validate scientific theories and models (Hancock and Diaz 2002; Meister 2000). Testing and evaluating these models are essential aspects of theory/model development process, allowing for the recognition of scientific advancements in the field. With this objective in mind, this study proposed a set of criteria for model evaluation in HFE and introduced a methodological procedure to apply these criteria to the case of trust in automation. The findings revealed differences between the two main classes of models. Conceptual models provide valuable insight into listings of variables that have or are assumed to have a direct causal effect on trust such as cultural variations, personality traits, and automation reliability. These models strive to consider all or the most significant elements that might have a causal impact on operators' trust and reliance on automation. However, testability and empirical validation of these models remain the biggest challenge to tackle. On the other hand, computational models incorporate mathematical representations that aim to predict or estimate levels of trust and can often be tested against data. Yet, these models can encompass only a limited number of causal factors and hence are less generalizable to various trust scenarios.

The analysis also indicated that there has been limited progress in TiA models over the years. This suggests that despite the efforts, the HFE community has struggled to significantly expand the frontiers of TiA research. The challenge lies in the complexity of trust as a psychological phenomenon and the inadequacy of the current modelling tools to effectively capture this complexity. The existing modelling approaches seem to be too simplistic and linear to effectively capture the intricate nature of trust in automation. Therefore, it is crucial for the HFE community to prioritize the adaptation of modelling approaches that can enhance our understanding of this phenomenon and, in turn, prove useful in real-world applications. Modelling approaches such as system dynamics, network dynamics, and agent-based modelling offer promising avenues for effectively modelling trust in automation. By leveraging these approaches, we may better grasp the complexity of trust by capturing interconnections and interactions among various entities in sociotechnical systems, emergent properties from these interactions, and the dynamic patterns of trust propagation and diffusion.

With regards to proposing an approach to evaluate HFE constructs, this study paves the way for new avenues of research. Firstly, although the proposed criteria are based on the known principles of the philosophy of science, further adjustments can be made to suit specific HFE models in future studies. Secondly, the rankings of the criteria based on the Best-Worst Method (BWM) reflect subjective assessments by researchers. Collecting and analysing judgments from Subject Matter Experts (SMEs) in future research can help reduce subjectivity. Similarly, achieving consensus among individual researchers on model ratings can enhance consistency in evaluation. Future studies may also consider matching some of

the criteria to the application. There may be no universal objective criteria weights. Matching the criteria to the target situation would allow individuals to select the right model for a particular situation, such as theory development or design.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Norges Forskningsråd.

References

Abbass, Hussein A., Eleni Petraki, Kathryn Merrick, John Harvey, and Michael Barlow. 2016. "Trusted Autonomy and Cognitive Cyber Symbiosis: Open Challenges." *Cognitive Computation* 8 (3): 385–408. https://doi.org/10.1007/s12559-015-9365-5

Abbass, Hussein A., Jason Scholz, and Darryn J. Reid. 2018. "Foundations of Trusted Autonomy: An Introduction." In *Foundations of Trusted Autonomy*, 1–12. Cham: Springer.

Adams, Barbara Dale, Lora E. Bruyn, Sébastien Houde, Angelopoulos, Paul. 2003. "*Trust in automated systems literature review*." Ministry of National Defence. Toronto, Canada: Defence Research and Development, Canada.

Akash, Kumar, Tahira Reid, and Neera Jain. 2018. "Adaptive Probabilistic Classification of Dynamic Processes: A Case Study on Human Trust in Automation." In 2018 Annual American Control Conference (ACC), 246–51. Milwaukee, WI: IEEE. https://doi.org/10.23919/ACC.2018.8431132

Akash, Kumar, Wan-Lin Hu, Tahira Reid, and Neera Jain. 2017. "Dynamic Modeling of Trust in Human-Machine Interactions." In 2017 American Control Conference (ACC), 1542–48. Seattle, WA, USA: IEEE. https://doi.org/10.23919/ACC.2017.7963172

Athey, Susan. 2017. "Beyond Prediction: Using Big Data for Policy Problems." *Science (New York, N.Y.)* 355 (6324): 483–485. https://doi.org/10.1126/science.aal4321

Azevedo, Jane. 1997. *Mapping Reality: An Evolutionary Realist Methodology for the Natural and Social Sciences*. New York, Albany: SUNY Press.

Azevedo-Sa, Hebert, Suresh Kumaar Jayaraman, Connor T. Esterwood, X. Jessie Yang, Lionel P. Robert, and Dawn M. Tilbury. 2021. "Real-Time Estimation of Drivers' Trust in Automated Driving Systems." *International Journal of Social Robotics* 13 (8): 1911–1927. https://doi.org/10.1007/s12369-020-00694-1

Bacharach, Samuel B. 1989. "Organizational Theories: Some Criteria for Evaluation." *The Academy of Management Review* 14 (4): 496–515. https://doi.org/10.2307/258555

Barber, Bernard. 1983. *The Logic and Limits of Trust*. New Brunswick, NJ: Rutgers University Press.

Baumeister, Roy F., and Brad J. Bushman. 2020. *Social Psychology and Human Nature*. Boston, MA, USA: Cengage Learning.

Bhakthavatsalam, Sindhuja, and Nancy Cartwright. 2017. "What's so Special about Empirical Adequacy?" *European Journal for Philosophy of Science* 7 (3): 445–465. https://doi.org/10.1007/s13194-017-0171-7

Billings, C. E. 1995. "Situation Awareness Measurement and Analysis: A Commentary." In *Proceedings of the International Conference on Experimental Analysis and Measurement of Situation Awareness*. Vol. 1. Daytona Beach, FL: Embry-Riddle Aeronautical University Press.

Bindewald, Jason M., Christina F. Rusnock, and Michael E. Miller. 2018. "Measuring Human Trust Behavior in Human-Machine Teams." In *Advances in Human Factors in Simulation and Modeling*, edited by Daniel N. Cassenti, 591:47–58. Advances in Intelligent Systems and Computing. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-60591-3_5

Blalock, Hubert M. 1969. *Theory Construction: From Verbal to Mathematical Formulations*. Prentice-Hall Englewood Cliffs, NJ.

Borsboom, Denny, Han L. J. van der Maas, Jonas Dalege, Rogier A. Kievit, and Brian D. Haig. 2021. "Theory Construction Methodology: A Practical Framework for Building Theories in Psychology." *Perspectives on Psychological Science : a Journal of the Association for Psychological Science* 16 (4): 756–766. https://doi.org/10.1177/1745691620969647

Burt, Ronald S., and Marc Knez. 1996. "Trust and Third-Party Gossip." *Trust in Organizations: Frontiers of Theory and Research* 68: 89.

Campbell, Donald T. 1988. *Methodology and Epistemology for Social Sciences: Selected Papers*. Chicago, IL: University of Chicago Press.

Caple, D. 2008. "Emerging Challenges to the Ergonomics Domain." *Ergonomics* 51 (1): 49–54. https://doi.org/10.1080/00140130701800985

Carnap, Rudolf. 1953. *Testability and Meaning*. New York: Appleton-Century-Crofts.

Carter, Stacy M., and Miles Little. 2007. "Justifying Knowledge, Justifying Method, Taking Action: Epistemologies, Methodologies, and Methods in Qualitative Research." *Qualitative Health Research* 17 (10): 1316–1328. https://doi.org/10.1177/1049732307306927

Cass, Elisa Maria. 2011. "Can Situation Awareness Be Predicted?: Investigating Relationships between CogScreen-AE and Pilot Situation Awareness." PhD Thesis, Carleton University.

Chen, Min., Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2018. "Planning with Trust for Human-Robot Collaboration." In Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, 307–315.

Chen, Min., Stefanos Nikolaidis, Harold Soh, David Hsu, and Siddhartha Srinivasa. 2020. "Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning." *ACM Transactions on Human-Robot Interaction* 9 (2): 1–23. https://doi.org/10.1145/3359616

Chien, Shih-Yi, Michael Lewis, Zhaleh Semnani-Azad, and Katia Sycara. 2014. "An Empirical Model of Cultural Factors on Trust in Automation." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58 (1): 859–863. https://doi.org/10.1177/1541931214581181

Chiou, Erin K., and John D. Lee. 2023. "Trusting Automation: Designing for Responsivity and Resilience." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 65 (1): 137–165. April, 001872082110099. https://doi.org/10.1177/00187208211009995

Cohen, Jacob, Patricia Cohen, Stephen G. West, and Leona S. Aiken. 1983. "Regression Applied Multiple Regression: Correlation Analysis for the Behavioral Sciences." *Correlation Analysis for the Behavioral Sciences*.

Cohen, Marvin S., R. A. J. A. Parasuraman, D. A. N. I. E. L. Serfaty, and R. Andes. 1997. *Trust in Decision Aids: A Model and a Training Strategy*." Arlington, VA: Cognitive Technologies, Inc.

Corbett, Martin. 2015. "From Law to Folklore: Work Stress and the Yerkes-Dodson Law." *Journal of Managerial Psychology* 30 (6): 741–752. https://doi.org/10.1108/JMP-03-2013-0085

Cramer, Kenneth M. 2013. "Six Criteria of a Viable Theory: Putting Reversal Theory to the Test." *Journal of Motivation, Emotion, and Personality: Reversal Theory Studies*. (February), 9–16. https://doi.org/10.12689/jmep.2013.102

De Keyser, V., F. Decortis, and A. Van Daele. 1988. "The Approach of Francophone Ergonomy: Studying New Technologies." *The Meaning of Work and Technological Options*. London: John Willey & Sons. *PMCid: PMC1050468*.

Degani, Asaf, and Michael Heymann. 2002. "Formal Verification of Human-Automation Interaction." *Human Factors* 44 (1): 28–43. https://doi.org/10.1518/0018720024494838

Dekker, Sidney, Paul Cilliers, and Jan-Hendrik Hofmeyr. 2011. "The Complexity of Failure: Implications of Complexity Theory for Safety Investigations." *Safety Science* 49 (6): 939–945. https://doi.org/10.1016/j.ssci.2011.01.008

Dekker, Sidney, and Erik Hollnagel. 2004. "Human Factors and Folk Models." *Cognition, Technology & Work* 6 (2): 79–86. https://doi.org/10.1007/s10111-003-0136-9

Desai, Munjal. 2012. "Modeling Trust to Improve Human-Robot Interaction." PhD Thesis, University of Massachusetts Lowell. https://www.yumpu.com/en/document/read/36708191/modeling-trust-to-improve-human-robot-interaction-umass-lowell-.

Desai, Munjal, Poornima Kaniarasu, Mikhail Medvedev, Aaron Steinfeld, and Holly Yanco. 2013. "Impact of Robot Failures and Feedback on Real-Time Trust." In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 251–58. IEEE. https://doi.org/10.1109/HRI.2013.6483596

Deutch, David. 1998. *The Fabric of Reality: The Science of Parallel Universes and Its Implications*." Viking Penguin: New York.

Deutsch, David. 2011. *The Beginning of Infinity: Explanations That Transform the World*. London, United Kingdom: Penguin UK.

Dienes, Zoltan. 2008. *Understanding Psychology as a Science: An Introduction to Scientific and Statistical Inference*. New York: Macmillan International Higher Education.

Douglas, Lisa, Denise Aleva, and Paul Havig. 2007. "Shared Displays: An Overview of Perceptual and Cognitive Issues." DivisionIn *12th International Command and Control Research and Technology Symposium: Adapting C2 to the 21st Century (Cognitive and Social Issues)*, 19–21. Newport, RI.

Drost, Ellen A. 2011. "Validity and Reliability in Social Science Research." *Education Research and Perspectives* 38 (1): 105–123.

Dubin, Robert. 1970. "Theory Building." *Philosophy and Phenomenological Research* 31 (2): 309. https://doi.org/10.2307/2105755

Dzindolet, Mary T., Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. "The Role of Trust in Automation Reliance." *International Journal of Human-Computer Studies* 58 (6): 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Earp, Brian D., and David Trafimow. 2015. "Replication, Falsification, and the Crisis of Confidence in Social Psychology." *Frontiers in Psychology* 6 (May): 621. https://doi.org/10.3389/fpsyg.2015.00621

Edwards, Ward, Harold Lindman, and Leonard J. Savage. 1963. "Bayesian Statistical Inference for Psychological Research." *Psychological Review* 70 (3): 193–242. https://doi.org/10.1037/h0044139

Endsley, Mica R. 2015. "Final Reflections: Situation Awareness Models and Measures." *Journal of Cognitive Engineering and Decision Making* 9 (1): 101–111. https://doi.org/10.1177/1555343415573911

Eronen, Markus I. 2021. "The Levels Problem in Psychopathology." *Psychological Medicine* 51 (6): 927–933. https://doi.org/10.1017/S0033291719002514

Fawcett, Jacqueline. 1986. "The Relationship of Theory and Research."

Fawcett, Jacqueline. 1988. "Conceptual Models and Theory Development." *Journal of Obstetric, Gynecologic & Neonatal Nursing* 17 (6): 400–403. https://doi.org/10.1111/j.1552-6909.1988.tb00465.x

Fawcett, Jacqueline. 2005. "Criteria for Evaluation of Theory." *Nursing Science Quarterly* 18 (2): 131–135. https://doi.org/10.1177/0894318405274823

Flach, John M. 1995. "Situation Awareness: Proceed with Caution." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37 (1): 149–157. https://doi.org/10.1518/001872095779049480

Frankfort-Nachmias, Chava, David Nachmias, and Jack DeWaard. 2014. *Research Methods in the Social Sciences*. Eighth edition. New York, NY: Worth Publishers.

Freedman, David A. 2010. *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*. New York: Cambridge University Press.

French, B., A. Duenser, Commonwealth Scientific and … A. Heathcote. 2018. "*Trust in Automation—a Literature Review*." *CSIRO Report EP184082)*. Canberra, Australia:.

Fried, Eiko I. 2020. "Theories and Models: What They Are, What They Are for, and What They Are About." *Psychological Inquiry* 31 (4): 336–344. https://doi.org/10.1080/1047840X.2020.1854011

Gambetta, Diego. 1988. "Trust: Making and Breaking Cooperative Relations."

Gao, F., A. S. Clare, J. C. Macbeth, and M. L. Cummings. 2013. "Modeling the Impact of Operator Trust on Performance in Multiple Robot Control." In *AAAI Spring Symposium - Technical Report*, SS-13-07:16–22. Palo Alto, CA. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84883414355&partnerID=40&md5=e368acd7ad55074c00ba454ea8eebff2.

Gao, Ji, and John D. Lee. 2006. "Extending the Decision Field Theory to Model Operators' Reliance on Automation in Supervisory Control Situations." *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 36 (5): 943–959. https://doi.org/10.1109/TSMCA.2005.855783

Garfinkel, Alan. 1982. "Forms of Explanation: Rethinking the Questions in Social Theory." *British Journal for the Philosophy of Science* 33 (4): 438–441.

Gauch, Hugh G. 2012. *Scientific Method in Brief*. Cambridge, UK: Cambridge University Press.

Getty, R. L. 1995. "Should We View Ergonomics as a Science, an Applied Engineering Practice or an Umbrella Multi-Discipline Program? What is Legitimate or Illegitimate Application of Ergonomics?." *Advances in Industrial Ergonomics and Safety VII*, London: Taylor & Francis.

Gould, Stephen. 1991. *Ever Since Darwin: Reflections in Natural History*. Penguin Books. https://books.google.com/books/about/Ever_Since_Darwin.html?hl=no&id=hb9k3LXnC6gC.

Guastello, Stephen J. 2001. "Nonlinear Dynamics in Psychology." *Discrete Dynamics in Nature and Society* 6 (1): 11–29. https://doi.org/10.1155/S1026022601000024

Guastello, Stephen J. 2017. "Nonlinear Dynamical Systems for Theory and Research in Ergonomics." *Ergonomics* 60 (2): 167–193. https://doi.org/10.1080/00140139.2016.1162851

Guo, Yaohui, and X. Jessie Yang. 2021. "Modeling and Predicting Trust Dynamics in Human–Robot Teaming: A Bayesian Inference Approach." *International Journal of Social Robotics* 13October. (8): 1899–1909. https://doi.org/10.1007/s12369-020-00703-3

Hancock, P. A., and D. D. Diaz. 2002. "Ergonomics as a Foundation for a Science of Purpose." *Theoretical Issues in Ergonomics Science* 3 (2): 115–123. https://doi.org/10.1080/14639220210123798

Hancock, Peter A., Deborah R. Billings, Kristin E. Schaefer, Jessie Y. C. Chen, Ewart J. de Visser, and Raja Parasuraman. 2011. "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction." *Human Factors* 53 (5): 517–527. https://doi.org/10.1177/0018720811417254

Harré, Rom. 1984. "Personal Being: A Theory for Individual Psychology." Oxford: Blackwell.

Hitchcock, Christopher, and James Woodward. 2003. "Explanatory Generalizations, Part II: Plumbing Explanatory Depth." *Nous* 37 (2): 181–199. https://doi.org/10.1111/1468-0068.00435

Ho, Geoffrey, Liana Maria Kiff, Tom Plocher, and Karen Zita Haigh. 2005. "A Model of Trust and Reliance of Automation Technology for Older Users." In AAAI Fall Symposium: Caring Machines, 45–50.

Hoff, Kevin Anthony, and Masooda Bashir. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust." *Human Factors* 57 (3): 407–434. https://doi.org/10.1177/0018720814547570

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. "Prediction and Explanation in Social Systems." *Science (New York, N.Y.)* 355 (6324): 486–488. https://doi.org/10.1126/science.aal3856

Hollnagel, Erik. 2002. "Time and Time Again." *Theoretical Issues in Ergonomics Science* 3 (2): 143–158. https://doi.org/10.1080/14639220210124111

Hoogendoorn, Mark, Syed Waqar Jaffry, Peter-Paul van Maanen, and Jan Treur. 2013. "Modelling Biased Human Trust Dynamics." *Web Intelligence and Agent Systems: An International Journal* 11 (1): 21–40. https://doi.org/10.3233/WIA-130260

Hou, Ming, Geoffrey Ho, and David Dunwoody. 2021. "IMPACTS: A Trust Model for Human-Autonomy Teaming." *Human-Intelligent Systems Integration* 3 (2): 79–97. https://doi.org/10.1007/s42454-020-00023-x

Howard, George S. 1985. "The Role of Values in the Science of Psychology." *American Psychologist* 40 (3): 255–265. https://doi.org/10.1037/0003-066X.40.3.255

Howson, Colin, and Peter Urbach. 1989. *ScientificReasoning: TheBayesian Approach*." La Salle, IL: Open Court Publishing Co.

Hu, Wan-Lin, Kumar Akash, Tahira Reid, and Neera Jain. 2019. "Computational Modeling of the Dynamics of Human Trust during Human–Machine Interactions." *IEEE Transactions on Human-Machine Systems* 49 (6): 485–497. https://doi.org/10.1109/THMS.2018.2874188

Huerta, Esperanza, TerryAnn Glandon, and Yanira Petrides. 2012. "Framing, Decision-Aid Systems, and Culture: Exploring Influences on Fraud Investigations." *International Journal of Accounting Information Systems* 13 (4): 316–333. https://doi.org/10.1016/j.accinf.2012.03.007

Hussein, Aya., Sondoss Elsawah, and Hussein Abbass. 2019. "A System Dynamics Model for Human Trust in Automation under Speed and Accuracy Requirements." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63 (1): 822–826. https://doi.org/10.1177/1071181319631167

Hussein, Aya., Sondoss Elsawah, and Hussein Abbass. 2020. "Towards Trust-Aware Human-Automation Interaction: An Overview of the Potential of Computational Trust Models." In https://doi.org/10.24251/HICSS.2020.047

Itoh, Makoto. 2011. "A Model of Trust in Automation: Why Humans over-Trust?." In SICE Annual Conference 2011, 198–201.

Jagacinski, Richard J., and John M. Flach. 2018. *Control Theory for Humans: Quantitative Approaches to Modeling Performance*. CRC press.

Jarvenpaa, Sirkka L., Kathleen Knoll, and Dorothy E. Leidner. 1998. "Is Anybody out There? Antecedents of Trust in Global Virtual Teams." *Journal of Management Information Systems* 14 (4): 29–64. https://doi.org/10.1080/07421222.1998.11518185

Jodlowski, Mark T. 2008. *Extending Long Term Working Memory Theory to Dynamic Domains: The Nature of Retrieval Structures in Situation Awareness*. Mississippi State University.

Jones, Debra G. 2015. "A Practical Perspective on the Utility of Situation Awareness." *Journal of Cognitive Engineering and Decision Making* 9 (1): 98–100. https://doi.org/10.1177/1555343414554804

Kaplan, Abraham. 1964. *The Conduct of Inquiry: Methodology for Behavioural Science*. Chandler Publishing.

Karwowski, Waldemar. 2005. "Ergonomics and Human Factors: The Paradigms for Science, Engineering, Design, Technology and Management of Human-Compatible Systems." *Ergonomics* 48 (5): 436–463. https://doi.org/10.1080/00140130400029167

Kelly, C., M. Boardman, P. Goillau, and E. Jeannot. 2001. "Principles and Guidelines for the Development of Trust in Future ATM Systems: A Literature Review." *European Organisation for the Safety of Air Navigation* 1 (1): 48.

Kennedy, William G. 2012. "Modelling Human Behaviour in Agent-Based Models." In *Agent-Based Models of Geographical Systems*, edited by Alison J. Heppenstall, Andrew T. Crooks, Linda M. See, and Michael Batty, 167–179. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-90-481-8927-4_9

Kerlinger, Fred N., and Howard B. Lee. 1986. *Foundations of Behavioral Research, Fort Worth*." TX*:* Holt, Rinehart, Winston.

Kerlinger, Fred Nichols. 1979. "Behavioral Research a Conceptual Approach."

Kivunja, Charles. 2018. "Distinguishing between Theory, Theoretical Framework, and Conceptual Framework: A Systematic Review of Lessons from the Field." *International Journal of Higher Education* 7 (6): 44. https://doi.org/10.5430/ijhe.v7n6p44

Kramer, Roderick M., Marilynn B. Brewer, and Benjamin A. Hanna. 1996. "Collective Trust and Collective Action." *Trust in Organizations: Frontiers of Theory and Research* 1 (1): 357–389.

Kraus, Johannes Maria. 2020. "Psychological Processes in the Formation and Calibration of Trust in Automation." PhD Thesis, Universität Ulm.

Kraus, Johannes, David Scholz, Dina Stiegemeier, and Martin Baumann. 2020. "The More You Know: Trust Dynamics and Calibration in Highly Automated Driving and the Effects of Take-Overs, System Malfunction, and System Transparency." *Human Factors* 62 (5): 718–736. https://doi.org/10.1177/0018720819853686

Krippendorff, Klaus. 2004. "Reliability in Content Analysis: Some Common Misconceptions and Recommendations." *Human Communication Research* 30 (3): 411–433. https://doi.org/10.1111/j.1468-2958.2004.tb00738.x

Krippendorff, Klaus. 2011. "Computing Krippendorff's Alpha-Reliability."

Kugler, Peter N., and Michael T. Turvey. 2015. *Information, Natural Law, and the Self-Assembly of Rhythmic Movement*. New York: Routledge.

Kuhn, Thomas S. 1962. *The Structure of Scientific Revolutions*. Chicago: Chicago (University of Chicago Press) 1962."

Kuhn, Thomas S. 1977. "Second Thoughts on Paradigms. The Essential Tension." *Selected Studies in Scientific Tradition and Change. TS Kuhn*. Chicago, Il/London, Chicago University Press.

Lakatos, Imre. 1970. "*Falsification and the Methodology of Scientific Research Programmes. Criticism and the Growth of Knowledge*." *I. Lakatos and A. Musgrave.*"." Cambridge, Cambridge University Press.
Lakatos, Imre. 1978. "Science and Pseudoscience." *Philosophical Papers* 1: 1–7.
Laudan, Larry. 1978. *Progress and Its Problems: Towards a Theory of Scientific Growth*. Vol. 282. California: Univ of California Press.
Laudan, Larry. 1986. "Science and Values." In *Science and Values*. California: University of California Press.
Lee, John, and Neville Moray. 1992. "Trust, Control Strategies and Allocation of Function in Human-Machine Systems." *Ergonomics* 35 (10): 1243–1270. https://doi.org/10.1080/00140139208967392
Lee, John, and Neville Moray. 1994. "Trust, Self-Confidence, and Operators' Adaptation to Automation." *International Journal of Human-Computer Studies* 40 (1): 153–184. https://doi.org/10.1006/ijhc.1994.1007
Lee, John D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
Lewicki, Roy J., and Barbara B. Bunker. 1996. "Developing and Maintaining Trust in Work Relationships." *Trust in Organizations: Frontiers of Theory and Research* 114: 139.
Lewicki, Roy J., Daniel J. McAllister, and Robert J. Bies. 1998. "Trust and Distrust: New Relationships and Realities." *The Academy of Management Review* 23 (3): 438–458. https://doi.org/10.2307/259288
Liang, Fuqi, Matteo Brunelli, and Jafar Rezaei. 2020. "Consistency Issues in the Best Worst Method: Measurements and Thresholds." *Omega* 96: 102175. https://doi.org/10.1016/j.omega.2019.102175
Lipton, Peter. 1990. "Contrastive Explanation." *Royal Institute of Philosophy Supplement* 27 (March): 247–266. https://doi.org/10.1017/S1358246100005130
Madhavan, Poornima, and Douglas A. Wiegmann. 2004. "A New Look at the Dynamics of Human-Automation Trust: Is Trust in Humans Comparable to Trust in Machines?." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 48:581–585. SAGE Publications Sage CA: Los Angeles, CA. https://doi.org/10.1177/154193120404800365
Madsen, Maria, and Shirley Gregor. 2000. "Measuring Human-Computer Trust." In *11th Australasian Conference on Information Systems*, 53:6–8. Gladstone, Australia: Citeseer.
Manstead, Antony Stephen Reid, and Andrew George Livingstone. 2008. "Research Methods in Social Psychology." *Introduction to Social Psychology: A European Perspective*: 20–40. (4th ed.). Oxford: Blackwell.
Marchionni, Caterina. 2012. "Geographical Economics and Its Neighbours—Forces towards and against Unification." In *Philosophy of Economics*, 425–458. Elsevier. Oxford, UK. https://doi.org/10.1016/B978-0-444-51676-3.50015-4
Mayer, Roger C., James H. Davis, and F. David Schoorman. 1995. "An Integrative Model of Organizational Trust." *The Academy of Management Review* 20 (3): 709–734. https://doi.org/10.2307/258792
McCloskey, Michael. 1983. "Intuitive Physics." *Scientific American* 248 (4): 122–130. https://doi.org/10.1038/scientificamerican0483-122
McElreath, Richard. 2018. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. UK: Chapman and Hall/CRC.
McKelvey, Bill. 2017. "Model-Centered Organization Science Epistemology." *The Blackwell Companion to Organizations* (6): 752–780.
Meehl, Paul E. 1967. "Theory-Testing in Psychology and Physics: A Methodological Paradox." *Philosophy of Science* 34 (2): 103–15.
Meehl, Paul E. 1978. "Theoretical Risks and Tabular Asterisks: Sir Karl, Sir Ronald, and the Slow Progress of Soft Psychology." *Journal of Consulting and Clinical Psychology* 46 (4): 806–834. https://doi.org/10.1037/0022-006X.46.4.806
Meister, David. 2000. "Theoretical Issues in General and Developmental Ergonomics." *Theoretical Issues in Ergonomics Science* 1 (1): 13–21. https://doi.org/10.1080/146392200308444
Meister, David. 2018. *The History of Human Factors and Ergonomics*. UK: CRC Press.

Meleis, A. I. 2012. *A Model for Evaluation of Theories: Description, Analysis, Critique, Testing, and Support*." Theoretical Nursing: Development and Progress, 179–206.
Merritt, Stephanie M., and Daniel R. Ilgen. 2008. "Not All Trust is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions." *Human Factors* 50 (2): 194–210. https://doi.org/10.1518/001872008X288574
Merton, Robert King. 1968. *Social Theory and Social Structure*. New York, NY: Simon and Schuster.
Moray, Neville, and T. Inagaki. 1999. "Laboratory Studies of Trust between Humans and Machines in Automated Systems." *Transactions of the Institute of Measurement and Control* 21 (4-5): 203–211. https://doi.org/10.1177/014233129902100408
Morgan, Mary S., and Margaret Morrison. 1999. *Models as Mediators*. Cambridge: Cambridge University Press Cambridge.
Morton, Adam. 1990. "Mathematical Modelling and Contrastive Explanation." *Canadian Journal of Philosophy Supplementary Volume* 16: 251–270. https://doi.org/10.1080/00455091.1990.10717228
Muir, Bonnie M. 1987. "Trust between Humans and Machines, and the Design of Decision Aids." *International Journal of Man-Machine Studies* 27 (5-6): 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5
Muir, Bonnie M. 1994. "Trust in Automation: Part I. Theoretical Issues in the Study of Trust and Human Intervention in Automated Systems." *Ergonomics* 37 (11): 1905–1922. https://doi.org/10.1080/00140139408964957
Muir, Bonnie M., and Neville Moray. 1996. "Trust in Automation. Part II. Experimental Studies of Trust and Human Intervention in a Process Control Simulation." *Ergonomics* 39 (3): 429–460. https://doi.org/10.1080/00140139608964474
Nam, Changjoo, Phillip Walker, Huao Li, Michael Lewis, and Katia Sycara. 2020. "Models of Trust in Human Control of Swarms with Varied Levels of Autonomy." *IEEE Transactions on Human-Machine Systems* 50 (3): 194–204. https://doi.org/10.1109/THMS.2019.2896845
Ngwenyama, Ojelanki. 2014. "Logical Foundations of Social Science Research." In *Advances in Research Methods for Information Systems Research*, 7–13. UK: Springer.
Nickerson, J. V., and R. R. Reilly. 2004. "A Model for Investigating the Effects of Machine Autonomy on Human Behavior." In 37th Annual Hawaii International Conference on System Sciences, 2004. Proceedings of The, 10 pp. Big Island, HI, USA: "Ieee." https://doi.org/10.1109/HICSS.2004.1265325
Niiniluoto, Ilkka. 1999. "Critical Scientific Realism."
Niiniluoto, Ilkka. 2017. "Optimistic Realism about Scientific Progress." *Synthese* 194 (9): 3291–3309. https://doi.org/10.1007/s11229-015-0974-z
Nomura, Tatsuya, Takayuki Kanda, Tomohiro Suzuki, and Kensuke Kato. 2008. "Prediction of Human Behavior in Human–Robot Interaction Using Psychological Scales for Anxiety and Negative Attitudes toward Robots." *IEEE Transactions on Robotics* 24 (2): 442–451. https://doi.org/10.1109/TRO.2007.914004
Øvergård, Kjell Ivar, Cato Alexander BjØrkli, and Thomas Hoff. 2008. "The Bodily Basis of Control in Technically Aided Movement." In *Spaces of Mobility*, 123–146. London: Routledge.
Parasuraman, Raja, and Victor Riley. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 39 (2): 230–253. https://doi.org/10.1518/001872097778543886
Parasuraman, Raja, Thomas B. Sheridan, and Christopher D. Wickens. 2008. "Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs." *Journal of Cognitive Engineering and Decision Making* 2 (2): 140–160. https://doi.org/10.1518/155534308X284417
Pearl, Judea. 2009. *Causality*. Cambridge: Cambridge University Press.
Péli, Gábor, and Michael Masuch. 1997. "The Logic of Propagation Strategies: Axiomatizing a Fragment of Organizational Ecology in First-Order Logic." *Organization Science* 8 (3): 310–331. https://doi.org/10.1287/orsc.8.3.310
Peterson, Sandra J., and Timothy S. Bredow, eds. 2013. *Middle Range Theories: Application to Nursing Research*. 3rd ed. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins Health.

Platt, John R. 1964. "Strong Inference: Certain Systematic Methods of Scientific Thinking May Produce Much More Rapid Progress than Others." *Science (New York, N.Y.)* 146 (3642): 347–353. https://doi.org/10.1126/science.146.3642.347

Popper, Karl. 1969. *Conjectures and Refutations: The Growth of Scientific Knowledge*. London: Routledge.

Popper, Karl. 1972. *The Logic of Scientific Discovery*. London: Hutchinson.

Prochaska, James O., Julie A. Wright, and Wayne F. Velicer. 2008. "Evaluating Theories of Health Behavior Change: A Hierarchy of Criteria Applied to the Transtheoretical Model." *Applied Psychology* 57 (4): 561–588. https://doi.org/10.1111/j.1464-0597.2008.00345.x

Ray, Marilyn A. 1990. "Critical Reflective Analysis of Parse's and Newman's Research Methodologies." *Nursing Science Quarterly* 3 (1): 44–46. https://doi.org/10.1177/089431849000300111

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

Rezaei, Jafar. 2015. "Best-Worst Multi-Criteria Decision-Making Method." *Omega* 53: 49–57. https://doi.org/10.1016/j.omega.2014.11.009

Rezaei, Jafar. 2016. "Best-Worst Multi-Criteria Decision-Making Method: Some Properties and a Linear Model." *Omega* 64: 126–130. https://doi.org/10.1016/j.omega.2015.12.001

Rezaei, Jafar. 2022. "BWM Solvers | Best Worst Method." 2022. https://bestworstmethod.com/software/.

Risjord, Mark. 2019. "Middle-Range Theories as Models: New Criteria for Analysis and Evaluation." *Nursing Philosophy : An International Journal for Healthcare Professionals* 20 (1): E 12225. https://doi.org/10.1111/nup.12225

Robinaugh, Donald J., Jonas M. B. Haslbeck, Oisín Ryan, Eiko I. Fried, and Lourens J. Waldorp. 2021. "Invisible Hands and Fine Calipers: A Call to Use Formal Theory as a Toolkit for Theory Construction." *Perspectives on Psychological Science : a Journal of the Association for Psychological Science* 16 (4): 725–743. https://doi.org/10.1177/1745691620974697

Rooij, Iris van, and Giosuè Baggio. 2021. "Theory before the Test: How to Build High-Verisimilitude Explanatory Theories in Psychological Science." *Perspectives on Psychological Science : a Journal of the Association for Psychological Science* 16 (4): 682–697. https://doi.org/10.1177/1745691620970604

Sadrfaridpour, Behzad, Hamed Saeidi, Jenny Burke, Kapil Madathil, and Yue Wang. 2016. "Modeling and Control of Trust in Human-Robot Collaborative Manufacturing." In *Robust Intelligence and Trust in Autonomous Systems*, edited by Ranjeev Mittu, Donald Sofge, Alan Wagner, and W.F. Lawless, 115–141. Boston, MA: Springer US. https://doi.org/10.1007/978-1-4899-7668-0_7

Salas, Eduardo. 2008. "At the Turn of the 21st Century: Reflections on Our Science." *Human Factors* 50 (3): 351–353. https://doi.org/10.1518/001872008X288402

Sanders, Mark S., and Ernest James McCormick. 1998. "Human Factors in Engineering and Design." *Industrial Robot: An International Journal*. 25 (2): 153–153. Emerald Group Publishing Limited: UK. https://doi.org/10.1108/ir.1998.25.2.153.2.

Sarter, Nadine B., and David D. Woods. 1991. "Situation Awareness: A Critical but Ill-Defined Phenomenon." *The International Journal of Aviation Psychology* 1 (1): 45–57. https://doi.org/10.1207/s15327108ijap0101_4

Saunders, M. N. K., Philip Lewis, and Adrian Thornhill. 2007. *Research Methods for Business Students*. 4th ed. Harlow, England ; New York: Financial Times/Prentice Hall.

Schaefer, Kristin. 2013. "The Perception and Measurement of Human-Robot Trust."

Seong, Younho, and Ann M. Bisantz. 2000. "Modeling Human Trust in Complex, Automated Systems Using a Lens Model Approach." *Automation Technology and Human Performance: Current Research and Trends* 1 (1): 95–100.

Shapiro, Lawrence. 2019. "A Tale of Two Explanatory Styles in Cognitive Psychology." *Theory & Psychology* 29 (5): 719–735. https://doi.org/10.1177/0959354319866921

Sheridan, Thomas B. 2019. "Extending Three Existing Models to Analysis of Trust in Automation: Signal Detection, Statistical Parameter Estimation, and Model-Based Control." *Human Factors* 61 (7): 1162–1170. https://doi.org/10.1177/0018720819829951

Silva, Mary C. 1986. "Research Testing Nursing Theory: State of the Art." *ANS. Advances in Nursing Science* 9 (1): 1–11. https://doi.org/10.1097/00012272-198610000-00003

Sind-Prunier, Paula. 1996. "Bridging the Research/Practice Gap: Human Factors Practitioners' Opportunity for Input to Define Research for the Rest of the Decade." In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 40 (17): 865–867. SAGE Publications Sage CA: Los Angeles, CA. https://doi.org/10.1177/154193129604001706

Solberg, Elizabeth, Magnhild Kaarstad, Maren H. Rø Eitrheim, Rossella Bisio, Kine Reegård, and Marten Bloch. 2022. "A Conceptual Model of Trust, Perceived Risk, and Reliance on AI Decision Aids." *Group & Organization Management* 47 (2): 187–222. https://doi.org/10.1177/10596011221081238

Stich, Stephen, and Shaun Nichols. 1992. "Folk Psychology: Simulation or Tacit Theory?" *Mind & Language* 7 (1-2): 35–71. https://doi.org/10.1111/j.1468-0017.1992.tb00196.x

Szalma, James L., and Grant S. Taylor. 2011. "Individual Differences in Response to Automation: The Five Factor Model of Personality." *Journal of Experimental Psychology. Applied* 17 (2): 71–96. https://doi.org/10.1037/a0024170

Thompson, J. M. T., H. B. Stewart, and Rick Turner. 1990. "Nonlinear Dynamics and Chaos." *Computers in Physics* 4 (5): 562–563. https://doi.org/10.1063/1.4822949

Trafimow, David. 2012. "The Role of Auxiliary Assumptions for the Validity of Manipulations and Measures." *Theory & Psychology* 22 (4): 486–498. https://doi.org/10.1177/0959354311429996

Van de, Ven, and H. Andrew. 2007. *Engaged Scholarship: A Guide for Organizational and Social Research*. Oxford: Oxford University Press on Demand.

Van Fraassen, Bas C. 1980. *The Scientific Image*. Oxford: Oxford University Press.

Van Gelder, Timothy, and Robert F. Port. 1995. "It's about Time: An Overview of the Dynamical Approach to Cognition." *Mind as Motion: Explorations in the Dynamics of Cognition* 1: 43.

Van Lange, Paul A. M. 2013. "What We Should Expect from Theories in Social Psychology: Truth, Abstraction, Progress, and Applicability as Standards (TAPAS)." *Personality and Social Psychology Review : An Official Journal of the Society for Personality and Social Psychology, Inc* 17 (1): 40–55. https://doi.org/10.1177/1088868312453088

Velicer, Wayne F., Geoff Cumming, Joseph L. Fava, Joseph S. Rossi, James O. Prochaska, and Janet Johnson. 2008. "Theory Testing Using Quantitative Predictions of Effect Size." *Applied Psychology = Psychologie Appliquee* 57 (4): 589–608. https://doi.org/10.1111/j.1464-0597.2008.00348.x

Vries, Peter de., Cees Midden, and Don Bouwhuis. 2003. "The Effects of Errors on System Trust, Self-Confidence, and the Allocation of Control in Route Planning." *International Journal of Human-Computer Studies, Trust and Technology* 58 (6): 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9

Wacker, John G. 2004. "A Theory of Formal Conceptual Definitions: Developing Theory-Building Measurement Instruments." *Journal of Operations Management* 22 (6): 629–650. https://doi.org/10.1016/j.jom.2004.08.002

Walker, Guy H., Neville A. Stanton, and Paul Salmon. 2016. "Trust in Vehicle Technology." *International Journal of Vehicle Design* 70 (2): 157. https://doi.org/10.1504/IJVD.2016.074419

Watts, Duncan J., Emorie D. Beck, Elisa Jayne Bienenstock, Jake Bowers, Aaron Frank, Anthony Grubesic, Jake M. Hofman, Julia M. Rohrer, and Matthew Salganik. 2018. "Explanation, Prediction, and Causality: Three Sides of the Same Coin?" UK: OSF Preprints." https://doi.org/10.31219/osf.io/u6vz5

Weick, Karl E. 1974. "Middle Range Theories of Social Systems." *Behavioral Science* 19 (6): 357–367. https://doi.org/10.1002/bs.3830190602

Wickens, Christopher D. 2008. "Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement." *Human Factors* 50 (3): 397–403. https://doi.org/10.1518/001872008X288420

Wilson, Jeanne M., Susan G. Straus, and Bill McEvily. 2006. "All in Due Time: The Development of Trust in Computer-Mediated and Face-to-Face Teams." *Organizational Behavior and Human Decision Processes* 99 (1): 16–33. https://doi.org/10.1016/j.obhdp.2005.08.001

Winsen, Roel van, and Sidney W. A. Dekker. 2015. "SA Anno 1995: A Commitment to the 17th Century." *Journal of Cognitive Engineering and Decision Making* 9 (1): 51–54. https://doi.org/10.1177/1555343414557035

Witkin, Stanley L., and Shimon Gottschalk. 1988. "Alternative Criteria for Theory Evaluation." *Social Service Review* 62 (2): 211–224. https://doi.org/10.1086/644543

Woodward, James. 2005. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford university press.

Xu, Anqi, and Gregory Dudek. 2012. "Trust-Driven Interactive Visual Navigation for Autonomous Robots." In 2012 IEEE International Conference on Robotics and Automation, 3922–29. https://doi.org/10.1109/ICRA.2012.6225171

Xu, Anqi, 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. IEEE.

Yarkoni, Tal, and Jacob Westfall. 2017. "Choosing Prediction over Explanation in Psychology: Lessons from Machine Learning." *Perspectives on Psychological Science : a Journal of the Association for Psychological Science* 12 (6): 1100–1122. https://doi.org/10.1177/1745691617693393

Ylikoski, Petri. 2007. "The Idea of Contrastive Explanandum." In *Rethinking Explanation*, 27–42. The Netherlands: Springer.

Article 2

Poornikoo M., Mansouri M. (2023), "Systems approach to modeling controversy in Human factors and ergonomics (HFE)," 18th Annual System of Systems Engineering Conference (SoSe), Lille, France, 2023, pp. 1-8,
https://doi.org/10.1109/SoSE59841.2023.10178634

Systems approach to modeling controversy in Human factors and ergonomics (HFE)

Mehdi Poornikoo
Department of Maritime Operations
University of South-Eastern Norway (USN)
Borre, Norway
mpo@usn.no

Mo Mansouri
Department of Science and Industrial Systems
University of South-Eastern Norway (USN)
Kongsberg, Norway
mo.mansouri@usn.no

2023 18th Annual System of Systems Engineering Conference (SoSe) | 979-8-3503-2723-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/SOSE59841.2023.10178634

***Abstract*— Contemporary debate regarding the shortcomings of human factors and ergonomics (HFE) models has inspired a growing interest in the HFE community to reconsider what constitutes a scientific model. Concerns are raised about the scientific credibility and adequacy of the existing models to explain human performance in sociotechnical systems. This study aims to address the debate through an epistemological understanding of different modeling approaches and discuss the suitability of each modeling approach for HFE problems in socio-technical systems. We argue that above anything, HFE is a systems discipline and therefore requires systems models and methodologies to support systems view on performance. The dominant outcome-oriented modeling approaches are ill-suited to study the complexity of today's sociotechnical systems. Alternatively, mathematical and simulation modeling can be beneficial in highlighting the inherent complexities, feedback loops, and temporal properties of HFE constructs.**

Keywords*— *Human factors and ergonomics, Human automation interaction, Cognitive system engineering, Systems thinking, folk models

I. INTRODUCTION

Nearly two decades ago, Dekker and Hollnagel [1] argued that HFE models and constructs and their relations with human performance may not be sufficiently credible but rather are "folk models". Several other scholars also raised concerns that constructs such as situational awareness, and trust in automation are theoretically vague, irrefutable, and overly generalizable [2]–[5]. By contrast, Parasuraman, Sheridan & Wickens [6] contend that HFE constructs received credibility through empirical research and therefore cannot be considered folk models. The debate was further discussed through "epistemological self-confidence" in HFE studies and whether a large body of evidence for a construct is a sign of concrete science or critical reasoning and skepticism about the constructs separates science from non-science [7].

Up to the present time, questions regarding the validity and credibility of HFE models are a subject of ongoing debates which necessitate a closer look into the core of the issue through understanding HFE as a scientific field, its modeling needs, and investigating whether the existing modeling approaches can satisfy the field's modeling aspirations. This paper intends to address the debate by exploring the modeling practices in HFE and suggest a way forward by adopting a systems approach to modeling.

II. FOLK MODEL CONTROVERSY IN HFE

Within cognitive sciences and philosophy of mind, the term "folk psychology" is denoted as "a collection of psychological principles and generalizations which… underlies our everyday explanation of behavior" [9, p.37]. One famous example of a folk model is the Yerkes-Dodson Law (YDL) developed by animal behaviorists more than 100 years ago. Yerkes and Dodson [9] performed a series of experiments to understand the speed of Japanese mice in distinguishing between black and white boxes under different levels of electric shock. The experimenters tested the association between electric shocks as stimulus strength and formation of habit i.e., learning speed. They found out that "moderate" shocks resulted in faster learning outcomes compared to "low" or "extreme" shocks. Additionally, they discovered that there is a linear relationship between variables; the higher the shock, the faster the learning outcome (up to the optimal point); making the results graphically look like an inverted-U curve.

Within the next 50 years, the original paper was only cited ten times in psychology journals [10]. However, by the 1970s, the YDL and inverted-U became the mainstream rule for the explanation of all psychological reasons [11], from human anxiety and task performance [12] to motivation and performance

979-8-3503-2723-6/23/$31.00 ©2023 IEEE

[13] and the contemporary interpretation of YDL as "some stress is necessary for optimal performance" [15, p. 359]. In many of the YDL applications in psychology, little or no empirical evidence was found to support the theory; yet such overgeneralization and explanation by substitution offered an incredible degree of immunity against falsification [11].

The longevity of folk models is attributed to offering a simple and seductively convincing way of describing human behavior and not because of their ability to predict or explain a phenomenon by decomposing its constructs into quantifiable units and defining the relationships among them [1], [15]. More importantly, the risk of incorrect folk models is their resistance to scientific checks that would highlight their shortcomings; and the fact that they can be generalized to inapplicable situations [16].

To address the folk model controversy, it is important to review the core content and structure of the HFE models. All models are also constrained by the validity of the assumptions that they ride on [17]. When the assumptions are true, theorems concerning the methods are valid. When the assumptions are false, the theorems do not hold. Therefore, it seems reasonable to investigate the nature of HFE sciences, and the epistemological assumptions of modeling practices in HFE, and finally to realize whether the HFE models and methods can help the field to advance by providing a useful and accurate account of the

III. Modeling Human behavior

It is widely accepted that one of the main aims of science is to provide an epistemologically sound account of the empirical world. Science certainly has other aims too, such as improving quality of life or producing useful technology, but these pragmatic aims are deferential to its core epistemological aim, which is to develop and deepen our empirical knowledge. Models are a crucial aspect of scientific inquiry because they offer a mechanism to comprehend phenomena and record that information in a way that can be shared with others. All models are abstractions since they exclude details that are assumed to be unimportant in favor of emphasizing the aspects of the phenomenon that are thought to be most important. Quite often, this selection is subjective and dependent on the modeler's preferences, however, this subjectivity has implications for the model's usefulness and/or accuracy [18].

Modeling human behavior in sociotechnical systems that are prone to variations of technical and contextual factors is a complex undertaking [19], where a reliable explanation of human behavior is reported to be nearly impossible [20], [21]. In particular, modeling human-automation interaction (HAI) has become the current modeling challenge in cognitive engineering. Hollnagel [22] discusses this issue and argues that while the use of technologies and computers in the 1980s was seen as the solution, it has now become a source of problems. Many systems have become too complex that the consequences of work situations are underspecified and hence partially unpredictable.

Primary modeling efforts in HFE were focused on creating models that show how a series of consequences may evolve from a certain set of possible causes [23], [24]. However, with the rise of sophisticated computing algorithms and the emergence of Big Data technologies, statistical machine learning (ML) models are now capable of predicting human behavior. Big data empowers ML algorithms to uncover complex patterns and make more timely and precise predictions than ever before [25]. On the other hand, the latest development in neuroscience and non-invasive tools for examining human brain activities (e.g., fMRI, EEG, MEG, Eye tracking) allows direct observation of underlying neural responses related to specific stimuli and provides useful data to divulge cognitive processes related to human behavior and decision making [26]. Nonetheless, ML and Big data are not able to reason '*how'* specific behavior occurs, and thus, their application in studying human behavior is limited to forecasting. This becomes exceptionally important since in applied fields such as HFE, the ultimate goal is to solve real-world problems through understanding the root causes of undesirable behavior, predicting future events based on the assumed causes, and subsequently, being able to modify the controllable causal variables in order to improve the overall system's performance.

The prevailing modeling practice in HFE involves models that are essentially listing variables that have or are believed to have a causal influence on a particular construct. Take trust in automation (TiA) as an example. It has been empirically reported that mood [27], personality traits [28], [29], self-confidence [30], and tendency to trust [28] influence trust in automation. In two separate meta-analysis studies, Hancock et al. [31] and Schaefer et al. [32] provided a comprehensive list of factors found to influence trust in automation. This style of modeling has led to a proliferation of several causal models for any given construct. The abundance of various models brings about questions regarding the model's validity and empirical adequacy. For instance, which of the models best represents the phenomenon? Is there a way to compare the relative worth of each model? Is it entirely possible to account for 'all' controllable variables in experimental settings involving human

entities? Such questions demand an understanding of the phenomenon being modeled and the way it is modeled. Before we dive into the modeling approaches, we briefly review the core and purpose of HFE as a scientific field to help us realize its modeling needs.

IV. HFE AS A SYSTEMS DISCIPLINE

HFE as we know it became recognized as a scientific field under the name of human factors engineering in the late 1940s. Prior to that, the intellectual work of Taylor [33] in Scientific Management theory was central to providing practical principles and techniques for analyzing and synthesizing workflow in order to enhance labor productivity. Although HFE has been commonly understood as a study of humans interacting with their environment, HFE is above anything, a systems discipline [34]–[37]. The system can be a work system such as a socio-technical or cognitive system, where the aim of HFE sciences is to simultaneously enhance performance and well-being through design processes and/or better integrating humans and the system [34]. Wilson [36] defines six overlapping characteristics of HFE as a systems discipline – systems focus, context, interactions, holism, emergence, and embedding. *Systems focus* emphasizes the importance of studying systems as a combination of interconnections among (in)organic materials, functions, processes, and ideas. *Context* refers to the idea that all behavior and performance occur in a setting or context. The context determines the system's boundaries and level of complexity. The context also highlights an important discussion regarding the validity of HFE laboratory studies because they are unable to account for the complexity and numerous factors in real-world settings [38], [39]. Even the most statistically significant results in the best controlled environment can only account for a slim fraction of the real variance in real settings [36]. *Interactions* imply that the nature of a system consists of parts interacting with each other. This view lies at the core of HFE methodologies and concepts, to understand and optimize the interactions between humans, technical systems, joint cognitive systems, etc. The notion of interaction is also tightly related to the systems' complexity of the system. *Holism* refers to the idea that HFE systems should be seen as a whole, meaning that physical, technical, cognitive, and social systems must be simultaneously studied in order to offer a viable solution for enhanced safety, or performance improvement. *Emergence* as the fifth feature of systems HFE is the acknowledgment of emergent properties of systems [40]. This is especially central to HFE as almost all systems in real use can portray characteristics not expected by the designers of the systems. The opposite scenario can also be true where the impact of poor design can be reduced by the ability of the user to find a way to make the system work despite its limitations. Thus, people may behave in creative and unpredicted ways which may be beneficial to the overall system performance. *Embedding* refers to the way that HFE practitioners operate which involves the inclusion of stakeholders and subject matter experts in the participatory ergonomics [41].

With these assumptions in mind, it is fair to argue that all HFE models can be seen as a network of constructs and systems because it is entirely possible to identify a set of interrelated elements that act in a closed-loop system which by definition is thinking in systems [42], [43]. By acknowledging the systems view of HFE, it is natural to review the epistemological assumptions underlying different modeling and methods in HFE to see whether the systems thinking can be reflected in the modeling practices or not.

V. TWO BASIC EPISTEMOLOGIES

Two fundamental epistemologies underlie the different methods that are necessary to study HFE problems involving humans and sociotechnical systems. Bruner [44] differentiates them based on two modes of thought; the paradigmatic or logico-scientific mode (variance) and the narrative (process) mode. The two modes provide a distinctive way of fabricating reality. While they are complementary to each other, they differ entirely in their method of validation. Aldrich [45] classifies the *'what'* and *'how'* questions in terms of outcome-driven and event-driven approaches.

A. Thinking in straight lines

In outcome-oriented thinking, everything can be explained by causal chains of events. From this perspective, the root causes are the events starting the chains of cause and effect, such as A and B as illustrated in Figure 1.

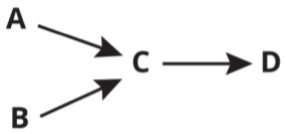


Figure 1. Outcome-oriented causal diagram

The outcome-driven approach to modeling, which Van de Ven [46] refers to as variance models intend to examine the relationship between independent and dependent variables to address the 'what' question,

e.g.; 'what are the antecedents and consequences associated with the phenomenon?'

With regard to causality, variance models require proof of co-variation, temporal precedence, and the absence of false association between the independent and dependent variables. Variance models use experimental and survey research designs, based on the general linear model that underlies most common statistical methods, such as ANOVA, regression, factor analysis, and structural equation modeling [46].

Poole et al. [47] highlight six assumptions underlying variance models.

1. The world constitutes fixed entities with varying attributes.
2. The basis of explanation is efficient causality.
3. The generality of explanation is dependent on the ability to apply it consistently across a wide range of situations.
4. The temporal sequence in which independent variables influence the dependent variable is irrelevant to the outcome.
5. Explanation should highlight immediate/direct causation.
6. Attributes have one and only one causal meaning over the course of time.

These assumptions signify a particular approach to constructing reality; a certain way of cutting up the world into researchable and measurable units. The variance modeling approach performs well for investigating research questions involving comparisons among entities or linear causal relationships between variables [48]. However, they are particularly limiting in studying social entities, particularly in sociotechnical systems.

B. Thinking in circles

"The reality is made up of circles, but we see straight lines" [39, p. 73].

In contrast to variance models and outcome-driven explanations, process models are event-driven explanations. The 'How' question explicates an observed series of events in terms of the underlying mechanisms that have the ability to cause events to occur in the real world and the specific conditions or contingencies under which these mechanisms operate [50].

Poole et al. [47] argue that the process approach views events as the most valid representation of what occurs in development and change processes with six contrasting assumptions to those of variance models.

1. The world is made up of entities that take part in events. These entities are not constant and may change over time.
2. The basis of explanation is final and formal causality, supplemented by efficient causality.
3. The generality of explanations is dependent on their flexibility.
4. The temporal sequence of events is extremely important.
5. Explanations should encompass layers of explanation ranging from immediate to distant.
6. An entity, attribute, or event may change in meaning over time.

These assumptions create a remarkable situation, where a particular cause may operate for only a limited time in a process model; that is, it never ceases to influence the entity as it forms part of the entity's history. More importantly, process models see the root causes as the forces emerging from particular feedback loops.

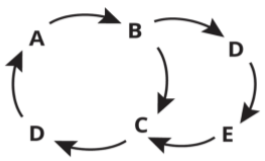


Figure 1. Event-driven causal diagram

Having addressed the difference between the two approaches, it is equally important to value their complementary relationship. Answers to a 'what' question typically presuppose or hypothesize an answer to a 'how' question. The logical reasoning behind a variance model - whether implicitly or explicitly- tells the story of how a series of circumstances led to an independent (input) variable having an impact on a dependent (outcome) variable. Therefore, carefully examining the process that is claimed to account for why an independent variable produces a dependent variable is one way to increase the robustness of answers to 'what' (variance theory) questions. Likewise, answers to 'how' questions are somehow pointless without an answer to the related variance theory questions of 'what caused it?' or 'what are its consequences?'

VI. Discussion and Concluding Remarks

For HFE as a systems discipline to progress, it is essential to realize the types of models that can be beneficial for studying human performance in sociotechnical systems. Analytical reduction in variance models is unable to explain *how* various entities and processes work together when multiple

impacts are present at the same time. Complexity is the defining feature of today's high-technology systems [51]. Interaction between a system's constituent parts can lead to complex behavior. Complexity invites us to pay attention to the system's relationships rather than the parts. These interactions produce properties of the system that cannot be obtained only by focusing on individual components. Complexity is a property of the system, not of its constituent parts [52]. The system's behavior cannot be boiled down to the behavior of its individual parts. Such systems must be examined as a whole if we are to study them. Furthermore, a complex system has irreversible circumstances. It is impossible to fully reconstruct the precise set of circumstances that led to the creation of a specific outcome. As linkages and interconnections change internally and adapt to their altering environments, complex systems undergo constant change. The system after the accident is not the same as the system before the accident because of the adaptive nature of complex systems [52]. Hence, the predictive power of variance models for retrospective analysis is very constrained [18]. Topmiller [53] argues that research in HFE systems poses a methodological challenge due to the complexity of the present systems, which demands simultaneous consideration of several interacting factors that influence different dimensions of both individual and group performance. The mainstream modeling in HFE however, follows the variance outcome-oriented approach, where HFE researchers and practitioners are given a range of methodological toolkits for studying aspects of individual operators, teams, and technical performance. These structured methods provide the basis for HFE discipline [54]. Examples can be found in methods and models of mental workload [55], situation awareness [56], trust in automation [57], and task analysis [58] which are frequently used within individual operator contexts. As pointed out, assumptions regarding variance models are ill-suited for the study of causality issues in social sciences. In particular, the notion of cause-effect relationships in sociotechnical systems seems to be nonlinear and bidirectional. Humans are continuously interacting within the environment they are in, consisting diverse range of entitles. While they impact other entities, they are simultaneously impacted.

Although in recent years HFE has witnessed few modeling attempts with systems thinking in mind, namely in accident analysis methods [59] and distributed situation awareness [60], the prevailing modeling practices still do not support systems view on performance [61], [62]. The problem becomes more apparent when researchers may mistakenly employ a variance model to study process questions. The downside of which is that the researcher is constrained to casting process dynamics into general linear relationships among variables. Compared to variance models, developing process models is relatively more complicated, but they can account for the complexity of events, feedback loops, temporal properties, and different time scales. Hence, they seem to be more suitable for studying social entities in HFE and Human Machine Interaction (HMI). More specifically, computational [63], [64] and dynamic simulation modeling [65]–[67] can be useful in providing a quantifiable language, specifying the relationships between the elements of HFE constructs, and addressing the temporal properties in dynamic social systems.

With that being said, and in line with the complementary connection between the two modeling approaches, future HFE process modelers can benefit from the existing variance models as there is a relatively good agreement regarding the causal factors for many constructs. The effort must be dedicated to addressing the inherent complexities in today's sociotechnical systems through the application of a systems approach and to provide enhanced causal models that can account for both 'what' and 'how' questions. This way, HFE, and cognitive system engineering can move toward a deeper and more meaningful understanding of human performance in social work settings and further formalizing HFE theories.

REFERENCES

[1] S. Dekker and E. Hollnagel, "Human factors and folk models," *Cognition, Technology & Work*, vol. 6, no. 2, pp. 79–86, May 2004, doi: 10.1007/s10111-003-0136-9.

[2] E. M. Cass, "Can situation awareness be predicted?: investigating relationships between CogScreen-AE and pilot situation awareness," PhD Thesis, Carleton University, 2011.

[3] L. Douglas, D. Aleva, and P. Havig, "Shared displays: an overview of perceptual and cognitive issues," AIR FORCE RESEARCH LAB WRIGHT-PATTERSON AFB OH WARFIGHTER INTERFACE DIVISION, 2007.

[4] J. M. Flach, "Situation Awareness: Proceed with Caution," *Hum Factors*, vol. 37, no. 1, pp. 149–157, Mar. 1995, doi: 10.1518/001872095779049480.

[5] M. T. Jodlowski, *Extending long term working memory theory to dynamic domains: The nature of retrieval structures in situation awareness*. Mississippi State University, 2008.

[6] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "Situation Awareness, Mental Workload, and Trust in Automation: Viable, Empirically Supported Cognitive Engineering Constructs," *Journal of Cognitive Engineering and Decision Making*, vol. 2, no. 2, pp. 140–160, Jun. 2008, doi: 10.1518/155534308X284417.

[7] S. W. A. Dekker, J. M. Nyce, R. van Winsen, and E. Henriqson, "Epistemological Self-Confidence in Human Factors Research," *Journal of Cognitive Engineering and Decision Making*, vol. 4, no. 1, pp. 27–38, Mar. 2010, doi: 10.1518/155534310X495573.

[8] S. Stich and S. Nichols, "Folk Psychology: Simulation or Tacit Theory?," *Mind & Language*, vol. 7, no. 1–2, pp. 35–71, Mar. 1992, doi: 10.1111/j.1468-0017.1992.tb00196.x.

[9] R. M. Yerkes and J. D. Dodson, "The relation of strength of stimulus to rapidity of habit-formation," *Punishment: Issues and experiments*, pp. 27–41, 1908.

[10] M. Corbett, "From law to folklore: work stress and the Yerkes-Dodson Law," *Journal of Managerial Psych*, vol. 30, no. 6, pp. 741–752, Aug. 2015, doi: 10.1108/JMP-03-2013-0085.

[11] K. H. Teigen, "Yerkes-Dodson: A law for all seasons," *Theory & Psychology*, vol. 4, no. 4, pp. 525–547, 1994.

[12] H. J. Eysenck, "A dynamic theory of anxiety and hysteria," *Journal of Mental Science*, vol. 101, no. 422, pp. 28–51, 1955.

[13] P. L. Broadhurst, "Emotionality and the Yerkes-Dodson law.," *Journal of experimental psychology*, vol. 54, no. 5, p. 345, 1957.

[14] L. A. Muse, S. G. Harris, and H. S. Feild, "Has the inverted-U theory of stress and job performance had a fair test?," *Human Performance*, vol. 16, no. 4, pp. 349–364, 2003.

[15] S. W. Dekker and D. D. Woods, "MABA-MABA or abracadabra? Progress on human–automation co-ordination," *Cognition, Technology & Work*, vol. 4, no. 4, pp. 240–244, 2002.

[16] M. Volkamer and K. Renaud, "Mental models–general introduction and review of their application to human-centred security," in *Number Theory and Cryptography*, Springer, 2013, pp. 255–280.

[17] D. A. Freedman, *Statistical models and causal inference: a dialogue with the social sciences*. Cambridge University Press, 2010.

[18] N. G. Leveson, "System safety engineering: Back to the future," *Massachusetts Institute of Technology*, 2002.

[19] K. I. Øvergård, C. A. Bjørkli, B. K. Røed, and T. Hoff, "Control strategies used by experienced marine navigators: observation of verbal conversations during navigation training," *Cognition, Technology & Work*, vol. 12, no. 3, pp. 163–179, Sep. 2010, doi: 10.1007/s10111-009-0132-9.

[20] E. Hollnagel, "Requirements for dynamic modelling of man-machine interaction," *Nuclear Engineering and Design*, vol. 144, no. 2, pp. 375–384, Oct. 1993, doi: 10.1016/0029-5493(93)90153-Z.

[21] E. Hollnagel and D. D. Woods, *Joint Cognitive Systems : Foundations of Cognitive Systems Engineering*. CRC Press, 2005. doi: 10.1201/9781420038194.

[22] E. Hollnagel, "Coping with complexity: past, present and future," *Cogn Tech Work*, vol. 14, no. 3, pp. 199–205, Sep. 2012, doi: 10.1007/s10111-011-0202-7.

[23] P. C. Cacciabue, "Understanding and modelling man-machine interaction," *Nuclear Engineering and Design*, vol. 165, no. 3, pp. 351–358, Sep. 1996, doi: 10.1016/0029-5493(96)01206-X.

[24] E. Hollnagel, "Requirements for dynamic modelling of man-machine interaction," *Nuclear Engineering and Design*, vol. 144, no. 2, pp. 375–384, Oct. 1993, doi: 10.1016/0029-5493(93)90153-Z.

[25] L. Zhou, S. Pan, J. Wang, and A. V. Vasilakos, "Machine learning on big data: Opportunities and challenges," *Neurocomputing*, vol. 237, pp. 350–361, May 2017, doi: 10.1016/j.neucom.2017.01.026.

[26] G. W. Lindsay, "Attention in Psychology, Neuroscience, and Machine Learning," *Front. Comput. Neurosci.*, vol. 14, p. 29, Apr. 2020, doi: 10.3389/fncom.2020.00029.

[27] S. M. Merritt, "Affective processes in human–automation interactions," *Human Factors*, vol. 53, no. 4, pp. 356–370, 2011.

[28] 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.

[29] J. L. Szalma and G. S. Taylor, "Individual differences in response to automation: The five factor model of personality," *Journal of Experimental Psychology: Applied*, vol. 17, no. 2, pp. 71–96, 2011, doi: 10.1037/a0024170.

[30] J. 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, Jan. 1994, doi: 10.1006/ijhc.1994.1007.

[31] P. A. Hancock, D. R. Billings, K. E. Schaefer, J. Y. C. Chen, E. J. de Visser, and R. Parasuraman, "A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction," *Hum Factors*, vol. 53, no. 5, pp. 517–527, Oct. 2011, doi: 10.1177/0018720811417254.

[32] K. E. Schaefer, J. Y. Chen, J. L. Szalma, and P. A. Hancock, "A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems," *Human factors*, vol. 58, no. 3, pp. 377–400, 2016.

[33] F. W. Taylor, "The Principles of Scientific Management," McMaster University Archive for the History of Economic Thought, 1911.

[34] J. Dul *et al.*, "A strategy for human factors/ergonomics: developing the discipline and profession," *Ergonomics*, vol. 55, no. 4, pp. 377–395, 2012.

[35] E. Hollnagel, "Human factors/ergonomics as a systems discipline? 'The human use of human beings' revisited," *Applied Ergonomics*, vol. 45, no. 1, pp. 40–44, Jan. 2014, doi: 10.1016/j.apergo.2013.03.024.

[36] J. R. Wilson, "Fundamentals of systems ergonomics/human factors," *Applied Ergonomics*, vol. 45, no. 1, pp. 5–13, Jan. 2014, doi: 10.1016/j.apergo.2013.03.021.

[37] G. H. Walker, P. M. Salmon, M. Bedinger, and N. A. Stanton, "Quantum ergonomics: shifting the paradigm of the systems agenda," *Ergonomics*, vol. 60, no. 2, pp. 157–166, Feb. 2017, doi: 10.1080/00140139.2016.1231840.

[38] M. Helander, *Design For Manufacturability: A Systems Approach To Concurrent Engineering In Ergonomics*. CRC Press, 1992.

[39] M. De Montmollin and L. Bainbridge, "Ergonomics and human factors," *Human Factors Society Bulletin*, vol. 28, no. 6, pp. 1–3, 1985.

[40] C. W. Johnson, "What are emergent properties and how do they affect the engineering of complex systems?," *Reliability Engineering and System Safety*, vol. 91, no. 12, pp. 1475–1481, 2006.

[41] H. Haines, J. R. Wilson, P. Vink, and E. Koningsveld, "Validating a framework for participatory ergonomics (the PEF)," *Ergonomics*, vol. 45, no. 4, pp. 309–327, 2002.

[42] A. D. Hall and R. E. Fagen, "Definition of system," in *Systems Research for Behavioral Sciencesystems Research*, Routledge, 2017, pp. 81–92.

[43] R. Carvajal, "Systemic-netfields: The systems' paradigm crisis. Part I," *Human Relations*, vol. 36, no. 3, pp. 227–245, 1983.

[44] E. M. Bruner, *The anthropology of experience*. University of Illinois Press, 1986.

[45] H. E. Aldrich, "Who wants to be an evolutionary theorist? Remarks on the occasion of the year 2000 OMT distinguished scholarly career award presentation," *Journal of Management Inquiry*, vol. 10, no. 2, pp. 115–127, 2001.

[46] A. H. Van de Ven, *Engaged scholarship: A guide for organizational and social research*. Oxford University Press on Demand, 2007.

[47] M. S. Poole, A. H. V. de Ven, K. Dooley, and M. E. Holmes, *Organizational Change and Innovation Processes: Theory and Methods for Research*. Oxford University Press, 2000.

[48] G. T. Payne, A. W. Pearson, and J. C. Carr, "Process and Variance Modeling: Linking Research Questions to Methods in Family Business Research," *Family Business Review*, vol. 30, no. 1, pp. 11–18, Mar. 2017, doi: 10.1177/0894486516679749.

[49] P. M. Senge, "The fifth discipline," *Measuring Business Excellence*, 1997.

[50] H. Tsoukas, "The validity of idiographic research explanations," *Academy of management review*, vol. 14, no. 4, pp. 551–561, 1989.

[51] C. Perrow, *Normal accidents: Living with high risk technologies*. Princeton university press, 1999.

[52] S. Dekker, P. Cilliers, and J.-H. Hofmeyr, "The complexity of failure: Implications of complexity theory for safety investigations," *Safety Science*, vol. 49, no. 6, pp. 939–945, Jul. 2011, doi: 10.1016/j.ssci.2011.01.008.

[53] D. A. Topmiller, "Methods: Past approaches, current trends and future requirements," *Manned Systems Design: Methods, Equipment, and Applications*, pp. 3–31, 1981.

[54] N. Stanton, P. M. Salmon, and L. A. Rafferty, *Human factors methods: a practical guide for engineering and design*. Ashgate Publishing, Ltd., 2013.

[55] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Advances in Psychology*, Elsevier, 1988, pp. 139–183. doi: 10.1016/S0166-4115(08)62386-9.

[56] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[57] 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, Nov. 1994, doi: 10.1080/00140139408964957.

[58] G. A. Klein, R. Calderwood, and A. Clinton-Cirocco, "Rapid decision making on the fire ground," in *Proceedings of the human factors society annual meeting*, Sage Publications Sage CA: Los Angeles, CA, 1986, pp. 576–580.

[59] N. Leveson, "A new accident model for engineering safer systems," *Safety Science*, vol. 42, no. 4, pp. 237–270, Apr. 2004, doi: 10.1016/S0925-7535(03)00047-X.

[60] N. A. Stanton *et al.*, "Distributed situation awareness in dynamic systems: theoretical development and application of an ergonomics methodology," *Ergonomics*, vol. 49, no. 12–13, pp. 1288–1311, Oct. 2006, doi: 10.1080/00140130600612762.

[61] P. M. Salmon, G. H. Walker, G. J. M. Read, N. Goode, and N. A. Stanton, "Fitting methods to paradigms: are ergonomics methods fit for systems thinking?," *Ergonomics*, vol. 60, no. 2, pp. 194–205, Feb. 2017, doi: 10.1080/00140139.2015.1103385.

[62] C. E. Siemieniuch and M. A. Sinclair, "Extending systems ergonomics thinking to accommodate the socio-technical issues of Systems of Systems," *Applied ergonomics*, vol. 45, no. 1, pp. 85–98, 2014.

[63] M. Itoh and K. Tanaka, "Mathematical modeling of trust in automation: Trust, distrust, and mistrust," in *Proceedings of the human factors and ergonomics society annual meeting*, SAGE Publications Sage CA: Los Angeles, CA, 2000, pp. 9–12.

[64] M. A. Sinclair and C. G. Drury, "On mathematical modelling in ergonomics," *Applied Ergonomics*, vol. 10, no. 4, pp. 225–234, Dec. 1979, doi: 10.1016/0003-6870(79)90215-1.

[65] P. C. Cacciabue, "Modelling and simulation of human behaviour for safety analysis and control of complex systems," *Safety Science*, vol. 28, no. 2, pp. 97–110, Mar. 1998, doi: 10.1016/S0925-7535(97)00079-9.

[66] M. C. Davis, H. Hughes, A. McKay, M. A. Robinson, C. N. van der Wal, and C. Natalie van der Wal, "Ergonomists as designers: computational modelling and simulation of complex socio-technical systems," *Ergonomics*, vol. 63, no. 8, pp. 938–951, Aug. 2020, doi: 10.1080/00140139.2019.1682186.

[67] R. Hanneman, *Computer-assisted theory building: modeling dynamic social systems*. Newbury Park, Calif: Sage Publications, 1988.

Article 3

Poornikoo, M., & Øvergård, K. I. (2022). Levels of automation in maritime autonomous surface ships (MASS): A fuzzy logic approach. Maritime Economics & Logistics, 24(2), 278-301. https://doi.org/10.1057/s41278-022-00215-z

Article 4

Poornikoo M., Gyldensten W., Vesin B., Øvergård, K. I. (In review) Trust in Automation (TiA): simulation model, and empirical findings in supervisory control of Maritime Autonomous Surface Ships (MASS), International Journal of Human-Computer Interaction

Trust in Automation (TiA): simulation model, and empirical findings in supervisory control of Maritime Autonomous Surface Ships (MASS)

Mehdi Poornikoo[1], William Gyldensten[1], Boban Vesin[2], Kjell Ivar Øvergård[3]

[1] Department of Maritime Operations, [2] Department of History, Business, and Social Sciences, [3] Department of Health, Social and Welfare Studies University of South-Eastern Norway (USN)

Abstract

Trust is recognized as a crucial element for effective interaction and utilization of autonomous systems. In supervisory control, trust determines the extent to which an operator relies on automation. This study introduces a dynamic simulation model for Trust in Automation (TiA) that encompasses trust development, deterioration, and recovery during interactions with automation. The model distinguishes itself from static conceptual models by offering a flexible, empirically testable framework, which can be tailored to fit a range of contexts. Utilizing a System Dynamics (SD) approach, the model reflects the non-linear and reciprocal nature of trust through dynamic feedback loops, generating behavioral patterns that align with empirical observations of trust evolution. An experimental study involving human participants in a simulated task with Maritime Autonomous Surface Ships (MASS) tests the model's empirical validity. The experiment focuses on behavioral responses to system malfunctions, emphasizing changes in monitoring strategies based on eye-tracking metrics. Consistent with the existing literature, the findings underscore the importance of initial conditions and aligning expectations with the system's performance to ensure the effective operation of autonomous systems.

Keywords: Trust in Automation, Dynamic modeling, Maritime Autonomous Surface Ships (MASS), Eye-tracking, System Dynamics

1 Introduction

The rapid progression of automation technologies, encompassing Maritime Autonomous Surface Ships (MASS), Autonomous Vehicles (AV), robotics, and autonomous web-based systems, is drastically transforming various facets of our daily lives. Understanding the dynamics of human interaction with these technologies is vital to harnessing their full potential. Trust emerges as a pivotal element shaping interactions between humans and automation (Kohn et al., 2021; Lee & Moray, 1992; Lee & See, 2004; Muir, 1994). Trust has been acknowledged as a crucial element for effective interaction and utilization of highly automated systems. As denoted by Lee and See (2004) and Porter et al. (2020), trust in automation also acts as a determinant of how much responsibility humans are willing to delegate to machines. Excessive trust may lead to over-reliance, where tasks are inappropriately assigned to automation, potentially resulting in neglectful oversight or failure to intervene when necessary. Conversely, distrust can lead to underutilization of automation capabilities, with operators unnecessarily intervening or disregarding the systems'

competencies (Parasuraman & Riley, 1997). Understanding and adjusting trust is paramount in enhancing human-automation interaction and integrating autonomous systems smoothly and effectively into daily operations (Gao et al., 2013; Lee & See, 2004).

The growing field of Trust in Automation (TiA) research has seen substantial progress in recent years. The diverse interest has generated numerous definitions, models, and measurements of trust. Comprehensive literature reviews (Basu & Singhal, 2016; Hancock et al., 2011, 2021; Madhavan & Wiegmann, 2007; Sanders et al., 2011; Schaefer et al., 2016; Tenhundfeld et al., 2022) have effectively synthesized research findings, and as a result, provide an ample understanding of the factors that may influence trust in automation. These works have also led to the development of diverse models to describe the development of trust in automation, offering distinctive perspectives on how trust is formed and evolves (Hoff & Bashir, 2015; Lee & See, 2004; Muir, 1994; Sheridan, 2019). One group of models stems from theoretical research with the aim of presenting a conceptual understanding of trust in automation. These models typically demonstrate multiple factors related to automation, individuals, and environmental characteristics, often depicted in network diagrams. Another set of studies involves computational models, striving to offer mathematical notations capable of predicting trust through the incorporation of causal factors and their interrelationships.

Despite the effort, ongoing TiA research faces five major challenges. *First*, many previous models have adopted a "snapshot" perspective on trust, often assessing trust at a single point in time. This static approach fails to fully recognize that trust is a dynamic phenomenon capable of both strengthening and declining due to moment-to-moment interaction with automation. With some exceptions (e.g., Hu et al., 2018; Lee & Moray, 1992; Lee & See, 2004; Xu & Dudek, 2015), there is limited understanding of trust formation, trust loss, and trust recovery, particularly in the context of autonomous systems (de Visser et al., 2020).

Second, While conceptual models (e.g., Hoff & Bashir, 2015; Lee & See, 2004) offer valuable insights into identifying the causal factors, the use of general terminology in these models limits their falsifiability and empirical validation (Dekker & Hollnagel, 2004; Poornikoo & Øvergård, 2023). A common limitation among many of these models is their lack of specificity regarding behavioral outcomes and the detailed patterns of trust evolution. This vagueness means that the models often fail to concretely predict the manner and direction in which trust will change in response to varying levels of the identified causal factors.

Third, Computational models such as the ones utilizing time series and regression formulations offer quantitative accounts of trust but primarily function as statistical models fitted to data (Hu et al., 2018; Lee & Moray, 1992, 1994; Moray et al., 2000; Muir & Moray, 1996). While these models provide a numerical understanding of trust variations and may successfully fit the observed data, their reliance on the specificities of the data often confines their broader applicability. Consequently, these models may not fully capture the nuances or generalize across various human-automation interaction contexts.

Fourth, Existing Trust in Automation (TiA) models frequently conceptualize the interactions between constructs and factors as unidirectional and linear pathways, adopting a simplistic stimulus-response framework. This approach significantly underrepresents the intricate interplay and coupling effects among human agents, automation technologies, and the operational environment (Jagacinski & Flach, 2018; Kugler & Turvey, 2015). Trust, as an outcome of prolonged dynamic interaction with automation on an infinite number of occasions, can be too complex to be adequately captured through a linear causal model. The inherent complexity of dynamic systems necessitates the acknowledgment of the temporal precedence of causal variables and the reciprocal and often simultaneous alterations in system properties (Guastello, 2017; Jagacinski & Flach, 2018; Van de Ven, 2007). For example, a system malfunction can have both an instant and a decaying reminiscence effect on future trust while trust itself may also vary non-linearly in response to changes in automation performance. These challenges highlight the need for careful consideration of model structure and the overall approach to modeling, as argued by Hollnagel (1993, 2002). Furthermore, nonlinear dynamic social systems may embody time delays for human perception and information processing (Jagacinski & Flach, 2018) or between communication errors and the effect these errors have on system performance (Øvergård et al., 2015), which can substantially affect the timing and formation of trust. Time delays can lead to oscillatory behavioral patterns (Sterman, 2000), further complicating the predictability and understanding of trust dynamics.

Fifth, Due to the limitations identified in existing Trust in Automation (TiA) models, their applicability and utility in real-world settings have been considerably constrained. Specifically, the existing models fall short in identifying the intervention points or areas for trust modification that could calibrate an operator's trust to accurately reflect the performance and capabilities of the automation system. As a result, the practical value of these models in guiding the engineering and design of automation systems remains largely unexplored and unestablished. The lack of specific, actionable insights from these models means that their contribution to enhancing the efficacy and safety of human-automation interaction in practical applications is yet to be fully realized and utilized.

To bridge these gaps, this paper proposes a system dynamic simulation model aimed at delineating the process of trust development, deterioration, and recovery through continuous interactions with automation. The model seeks to address the deficiencies of previous models and provides three key advantages. *Firstly*, it aligns closely with the existing trust literature and empirical studies, enhancing the model's explicability and generalizability across various contexts. *Secondly,* the model holds a high level of adaptability, capable of simulating a wide array of trust dynamics tailored to different scenarios and conditions. This flexibility allows it to reflect diverse real-world situations and potential changes in automation and user behavior. *Lastly,* the model emphasizes the role of model structure in determining the behavioral outcomes of trust interactions. By focusing on the structure, it moves towards a deeper understanding of the underlying mechanisms and patterns of trust formation and erosion, thereby contributing to more effective design and management of human-automation systems.

In addition, this study uses an experiment with a within-subject design, wherein 30 human participants engaged in a simulated task involving the monitoring of an unmanned autonomous vessel. This empirical facet is designed to test a segment of the proposed model where a shorter time scale allows for laboratory testing, more specifically in trust response to a system malfunction. The objective is to assess the model's predictions and to observe the corresponding behavioral reactions of the operators. Key focus areas include examining changes in the participants' visual scanning patterns and monitoring strategies in response to the malfunction. This approach aims to provide concrete insights into how operators interact with and adjust to automated systems under varying conditions, thereby reinforcing the model's relevance and applicability to human-automation interaction scenarios.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature on trust in automation. Section 3 formulates a system dynamics model of trust in automation and discusses its feedback loop structure and behavioral outcomes. Section 4 presents an empirical study and examines the results. Section 5 discusses the findings, and Section 6 concludes the study and suggests future research.

2 Theoretical backgrounds

A widely used definition of human-automation trust, put forth by Lee and See, characterizes trust as "*the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability*." (2004, p. 51). In this definition, the central idea remains that trust revolves around the belief that automation will effectively perform a task in a context marked by uncertainty and vulnerability. This perspective aligns with the concept of trust in interpersonal relationships, where uncertainty and vulnerability are integral components. Another definition defines trust as "*an attitude of confident expectation in an online situation of risk that one's vulnerabilities will not be exploited.*" (Corritore et al., 2003, p. 740). This definition is in line with Mayer, Davis, and Schoorman's description of 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*" (Mayer et al., 1995, p. 712). This definition emphasizes that expectation is a crucial element in the trust relationship, which is also pointed out by several other scholars (e.g., De Vries et al., 2003; Muir, 1987). Among various definitions of trust, this study adopts the definition proposed by Mayer et al. (1995). This concise definition will serve as our guiding framework for later developing a model of trust in automation.

Rempel et al. (1985) suggested that trust is a dynamic attitude shaped by specific dimensions that gradually form over time. They identified predictability, dependability, and faith as the three key dimensions that impact an individual's willingness to accept a trustee, laying the foundation for trust. This concept was later integrated into many early models of trust in the

automation (e.g., Lee & Moray, 1992, 1994; Lee & See, 2004; Muir, 1994; Muir & Moray, 1996).

Lee and Moray (1992) presented a first step toward modeling trust in automation and employed linear regression models to investigate the factors influencing trust in automation. In their subsequent work, Lee and Moray (1994) developed and tested several linear models to identify factors affecting trust. These models indicated that self-reported trust was influenced by two main factors: the system performance and the occurrence of a system fault. To account for the memory of trust and the impact of past performance, Lee and Moray (1992, 1994) introduced the Autoregressive Moving Average Vector (ARMAV) model, a stochastic time series model for trust. The utility of this trust model is somewhat constrained by its specificity and the context for which it was designed. It necessitates input data about the performance metrics and fault occurrences of the automation system involved in the particular experiment. This requirement means that the model's applicability may be limited to scenarios like the one it was originally developed for. Additionally, the model does not offer an underlying rationale for conceptualizing trust as a linear combination of prior trust levels, performance metrics, and fault incidences.

Muir (1994) integrated the trust models proposed by Barber (1983) and Rempel et al. (1985) and developed a linear model. According to Muir's model, trust can be predicted as an outcome of the expectation of persistence (P), technically competent performance (TCP), and fiduciary responsibility (FR). The trust model was further extended as a linear regression of predictability, dependability, faith, competence, responsibility, and reliability. A notable limitation of the mode remains as to how to operationalize its components such as dependability and faith (Rodriguez Rodriguez et al., 2023).

Gao and Lee (2006) proposed the Extended Decision Field Theory (EDFT) which is based on a dynamic-cognitive approach to human decision-making in order to describe “preference dynamics”. They employed an autoregressive approach that considers the linear combination of the previous preference and new input on the current preference in an uncertain environment. This model has been utilized to create a quantitative account of trust and self-confidence, linked to decision-making in automation usage. It incorporates the construction of belief in the automation’s capability or the operator’s manual capability using a piece-wise function. The EDFT model was able to successfully replicate empirical findings related to the inertia of trust and the relationship between trust, self-confidence, and reliance.
Xu and Dudek (2015) presumed a connection between the level of reliance and the level of trust, suggesting that reliance can serve as an indicator of trust. To delve into this concept, they developed the Online Probabilistic Trust Inference Model (OPTIMO), a Dynamic Bayesian Network model designed to capture a person's level of trust in a robot teammate.
Akash et al. (2017) introduced a three-state model of trust in automation that can account for biases in human behavior arising from perceptions of past trust and expectations. This model assumes that the change in trust, denoted as $T(n+1) - T(n)$, is linearly dependent on three factors: (1) the difference between experience and present trust ($E(n) - T(n)$), (2) the difference between cumulative trust and present trust ($CT(n) - T(n)$), and (3) the

difference between expectation bias and present trust ($BX(n) - T(n)$). If the present experience is less than the present trust level, the predicted trust level decreases, and vice versa. The model faces challenges in terms of its applicability to a broader range of scenarios, primarily due to its reliance on directly querying participants about their trust in automation. This approach presupposes the ability to measure current trust levels as a basis for forecasting future trust (Rodriguez Rodriguez et al., 2023).

Yang et al. (2017) employed a first-order linear time-invariant dynamic system and discovered that the average trust in automation stabilizes through recurring interactions. Recently, Guo and Yang (2021) employed a Beta distribution to construct a personalized trust prediction model, applying Bayesian inference to compute the Beta distribution parameters. The model is based on certain assumptions including trust at time t is influenced by trust at time $t-1$, negative experiences with automation have a more substantial impact on trust than positive experiences, and trust in automation stabilizes after multiple interactions. The authors acknowledged some limitations of the model such as the assumptions that the automation's ability remains constant across all interactions and the automation's performance is either consistently good or bad. Despite these limitations, the model, utilizing automation's performance and reliability alongside the human's self-reported trust history, exhibits potential for generalization to other human-automation task scenarios capturing changes in these measures.

Lewis and Weigert (2012) highlight the importance of capturing feedback loops in trust development, asserting that trust relationships have histories. Jonker and Treur (1999) delve into the dynamics of trust, describing it as the "evolution of trust over time." They offer a modeling framework that investigates the reciprocity of trust and the development and deterioration of trust between agents from an experiential perspective. Manzey et al. (2012) identified two feedback loops in the process of human trust adjustment: a positive and a negative feedback loop. The positive loop is activated by instances of automation success, while the negative loop is triggered by instances of automation failure. It is noteworthy that the negative feedback loop has a more pronounced impact on trust adjustment compared to the positive feedback loop (Yang et al., 2016).

Among conceptual models, the dynamic model of trust and reliance proposed by Lee and See (2004) presents a fundamental theoretical framework. At the heart of this closed-loop system, human operators acquire information about the physical state of the system through a display, and based on this information, they construct their own beliefs and assessments regarding the system's current state. The level of trust is then formed based on these beliefs about the automation's capability and the current state of the system. Depending on their level of trust, the operator may decide (*i.e.,* intend) whether to use the system or intervene as necessary. The action taken by the operator directly influences the state of the automation.
Recently, Sheridan (2019) extended Lee and See's (2004) model and mapped the model's elements to a Kalman system of human control (See Figure 1) including the automation (physical reality of cause-effect), display (measurement of result of action), information analysis and belief formation (discrepancy in estimation of state), trust evolution (internal or

mental model of reality), intention formation (state-based policy deciding action), and reliance action (physical action to modify state). The Kalman modification produces two intermediate feedback loops. The first feedback mechanism is responsible for measuring the discrepancy between what the actual automation's displayed information conveys and what the internal model (i.e., trust evolution) expects. This discrepancy between the two forms the basis for adjusting the internal trust evolution model. In other words, it allows the trust model to adapt and better align with the observed reality by modifying the internal trust belief based on this comparison. The second feedback loop enables the internal mental model of the actual automation to "anticipate" how a new action will modify the automation state. It provides the operator with the ability to foresee the consequences of their actions, allowing for more informed and proactive decision-making.

To illustrate the model's applicability to human behavior, Sheridan (2019) exemplifies walking down a flight of stairs with your arms full. In this situation, you cannot directly see your feet, making the placement of each foot on the next step a partially open-loop extrapolation estimate or "dead reckoning" based on your trust in your ability to navigate the stairs safely. The loop allows for the updating of your internal trust model to mediate how far you trust in each step of the process. However, even in this scenario, there's a need for real-world confirmation or closed-loop validation, which requires feedback on the discrepancy between where you think your feet are (internal trust model) and where they actually are (actual foot placement represented). The feedback mechanism, which compares the anticipated actions to the real-world outcomes, ensures that your trust in the process remains grounded in reality and allows for adaptive adjustments based on real-world conditions.

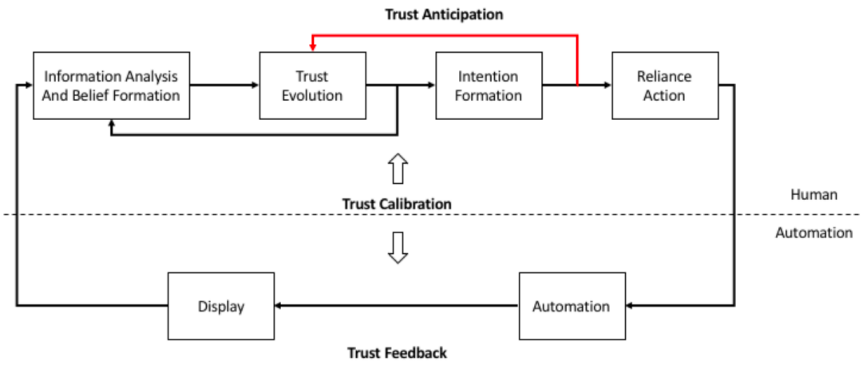


Figure 1. Modified Lee and See's (2004) model of trust in automation (Sheridan, 2019)

One important factor to consider in the evolution of trust is the time delays in the human perception and processing of information (Boubin et al., 2017; Lee & See, 2004). Unlike classical control systems, where these time lags tend to be relatively short, it may take human operators significantly longer to accurately perceive displayed information and form trust in automation. The human operator makes decisions based on their internal belief and the anticipated actions toward the automation system. However, the feedback from the system can arrive after some time lag. The significance of this time lag becomes more pronounced as the gap between the internal belief and the feedback from the actual system increases. In other words, the longer the time lag, the less the operator can expect their internal belief and

the actual feedback to closely align. Dealing with time delays or long inertial time constants in dynamic systems is something that operators experience in various scenarios. For example, consider a shore-based operator navigating a large ship making open-loop movements, trusting that these actions will lead to the desired results. However, verification of the outcome only occurs after a time delay due to the slow response of the system, with the hope that the system will not overshoot and produce disproportionate results. This process is described as an iteration of "trust-action-verify". The operator trusts the system, acts, and then waits for verification (Sheridan, 2019).

The time delays inherent in human-automation interactions necessitate a trust model that accommodates both open-loop decision-making and closed-loop verification based on real-time feedback, to ensure an adaptable trust in automation.

More importantly, the implementation of the assumptions in Lee & See (2004) and Sheridan (2019) trust models would require computer systems to simulate the interaction of the various 'blocks' within the model. These blocks may include perception and expectation of the system, belief formation, intention, or decision-making. The level of realism in the simulation would depend on the specific research or application context. For instance, the display of information in Lee and See's (2004) framework could be modeled to represent perfect information transfer, or it could introduce noise and uncertainties to simulate real-world factors like inattention or system limitations (Sheridan, 2019). This flexibility allows researchers to tailor the model to the specific conditions they want to investigate.

3 Proposed Model of Trust in Automation

This section introduces a simulation model designed to depict the evolving nature of trust in automation, stemming from continuous interactions with automated systems. The model aims to capture characteristics inherent to this dynamic, including time delays and nonlinear responses, thereby providing a comprehensive understanding of trust dynamics within an autonomous context.

3.1 Method

This study utilizes the principles of System Dynamics (SD) to develop a continuous event simulation model of trust in automation. System dynamics (SD) is a computer-based modeling approach that employs simulation to gain insights into the dynamics of complex systems over time (Sterman, 2000). The primary objective of system dynamics is to comprehend how a system's behavior emerges and to leverage this understanding to explore how changes within the structure of that system can influence its behavior. As a modeling and simulation approach, system dynamics excels at describing complex, non-linear, and often counter-intuitive behaviors driven by feedback mechanisms. It is especially suitable for systems characterized by feedback relationships and time delays (Azar, 2012; Lane & Schwaninger, 2008).

With roots in non-linear control theory (Forrester, 1987, 1997; Towill, 1993), SD is a versatile method that finds applications in diverse settings, ranging from highly practical contexts (e.g., Ghaffarzadegan et al., 2017; McCarthy et al., 2014; Williams, 1997) to more

theoretical domains (e.g., Gambardella et al., 2017; A. Sastry, 1998; S. Sastry, 2013; Sterman & Wittenberg, 1999; Wittenberg, 1992). In theoretical SD research, the primary outcomes are the development of new theories, or the adaptation and refutation of existing ones related to dynamic phenomena. SD employs both causal tracing, a potent technique for capturing mental models, and differential equations to represent the temporal changes within systems. System Dynamics models significantly enhance falsifiability, allowing for both logical- and empirical testing of each interrelationship within the model (Lane & Schwaninger, 2008; Schwaninger & Grösser, 2008).

As proposed by Lane (2000), the theory of System Dynamics is primarily a structural theory, distinguishing it from content theories. This posits that the developmental patterns of social systems over time are explicable through endogenous processes, which are represented by feedback loops, rates, and stock variables.

In contrast to other modeling approaches that primarily rely on data to describe and predict system behavior, SD revolves around the creation of formal models that capture dynamic patterns as continuous feedback systems. These models incorporate hypotheses about the causal relationships among parameters and variables as functional components, considering the results of their interactions. In this context, loops become higher-level units of analysis, each with a distinct purpose and varying degrees of significance over time. A single variable may belong to multiple feedback loops. Rather than examining the potential causal connections between individual pairs of variables, system dynamics modelers take a step back to understand the broader causative structures.

There are several reasons why SD is particularly well-suited for the modeling of human automation interaction. Firstly, given that human performance typically does not conform to linear models (Gao & Lee, 2006), incorporating non-linear behavioral and performance representations becomes necessary to ensure the model's external validity (Sweetser, 1999). Secondly, SD models offer the flexibility to incorporate both qualitative and quantitative data (Sterman 2000), with qualitative data often being indispensable for human performance modeling (Hancock & Szalma, 2004). Thirdly, SD models prove to be highly effective in depicting the effects of latencies and feedback interactions within the system, a vital aspect of modeling a human operator and assessing the influence of delays in perceiving system performance on operator behavior and trust (Cummings & Clare, 2015).

3.2 Model components

Literature suggests three primary sources that influence trust in automation: (1) the characteristics of the automation itself, (2) the trustor (or user), and (3) the environmental context in which interaction occurs (Hancock et al., 2011; Hoff & Bashir, 2015; Schaefer et al., 2016). Within each of these categories, several specific factors can introduce variability and complexity to trust dynamics. However, to construct a coherent and manageable model of trust dynamics, it is often practical to start with only a handful of these factors deemed most critical. While incorporating additional factors might enhance the model's robustness

and predictive accuracy, the fundamental behavioral patterns of trust can typically be captured with a few essential components. This approach allows for a more focused and tractable analysis, providing a balanced perspective between complexity and understandability in modeling trust dynamics.

A. System performance (reliability)

Studies have consistently shown a strong relationship between trust and the reliability of the automation (Moray & Inagaki, 1999; Parasuraman & Manzey, 2010; Parasuraman & Riley, 1997; Riley, 1994). When system reliability declines, trust, and trust expectations tend to systematically decline (Moray et al., 2000). System reliability pertains to automation with some error rate (Lewis et al., 2018).

B. System malfunctions

Although related to system reliability, system malfunctions must be considered separately due to their discrete occurrences. System malfunctions refer to "*sudden, unpredicted errors related to a system's reliability within its area of application*" (Kraus et al., 2020, p. 1) that can lead to disruption of normal operation and result in a fallback strategy (Emzivat et al., 2017). System malfunctions are usually singular events and can be studied as a "shock" in the automation's continuous performance. Trust can be affected differently depending on the type and magnitude of the malfunction. In cases of mild and temporary faults, trust may decline briefly and then recover. However, when a failure impairs the automation's capabilities, trust may decline continuously until operators decide to rebuild it (Itoh et al., 1999; Lee & Moray, 1992). A study conducted by Lee & Moray (1992) revealed that when persistent system malfunctions were present, trust in automation reached its lowest point after six trials but trust gradually recovered even as faults persisted. The dynamic changes in trust resulting from system malfunctions may not only happen instantly but evolve over time (Lee & Moray, 1992). A time-series analysis demonstrated that the impact of malfunction on trust can be modeled using a first-order differential equation, where it is suggested that the most substantial effect is observed immediately after a failure, with a residual effect extending over time (Lee and Moray, 1994). Muir & Moray (1996) revealed that malfunctions that are erratic in magnitude reduced trust more than faults that were large and continuous.

While system malfunctions can impact trust in automation, this effect is particularly pronounced when individuals lack prior knowledge of these malfunctions. Studies have indicated that when people are aware of potential malfunctions in automation, these issues do not necessarily erode their trust in the system (Riley, 1994). One plausible explanation for this phenomenon is that foreknowledge of potential automation malfunctions reduces uncertainty and the associated risks tied to using the system.

C. Individual's characteristics

Trust in automation is not solely affected by the attributes of the automation but is significantly influenced by a person's subjective perception of these attributes (Lee & See, 2004; Merritt & Ilgen, 2008). Merritt and Ilgen (2008) discovered that an individual's perception of automation is influenced not only by the actual characteristics of the machine

but also by the person's inherent propensity to trust machines. According to Mayer et al. (1995), trust propensity is a stable trait that varies from person to person, depending on factors such as developmental experiences, personality type, and cultural background. Demographic factors such as age and gender, have been the subject of research regarding their relationship with trust in automation. While some studies have shown age-related differences (Donmez et al., 2006), these differences appear to be context-dependent (Hoff & Bashir, 2015). More recently, Hartwich et al. (2019) investigated the impact of age groups on trust in an automated driving system and found no significant differences among the groups. Similarly, there is currently no consensus on the effects of gender on trust in automation (Hoff & Bashir, 2015). In the realm of general personality traits, research has frequently explored the traits of the Five-Factor Model of Personality (John & Srivastava, 1999; McCrae & John, 1992) concerning trust. Notably, extraversion has been consistently linked to higher levels of interpersonal trust (Evans & Revelle, 2008), and this positive association extends to the propensity to trust in automation (Merritt & Ilgen, 2008). Conversely, neuroticism is negatively related to interpersonal trust (Evans & Revelle, 2008), and while there is no direct evidence of such a link to trust in automation, a connection between neuroticism and accepting recommendations by automated systems suggests a potential relationship (Szalma & Taylor, 2011). Additionally, agreeableness and conscientiousness have been positively associated with individuals' initial trust automation (Chien et al., 2016).

D. Environmental characteristics

Regarding environmental-related factors, studies indicate that the extent to which individuals rely on automation is influenced by the level of risk associated with their decision to use it (Riley, 1996). When the likelihood of negative outcomes is higher, individuals tend to be more hesitant to use automation, and once their trust is diminished, it takes them longer to regain trust in high-risk situations compared to low-risk scenarios (Riley, 1994).

3.3 Model Formulation

A. Model assumptions

Trust as a dynamic phenomenon (Kim et al., 2009; Muir & Moray, 1996) typically unfolds in three phases: trust formation, where trustors decide to trust trustees and may gradually increase their trust; trust dissolution, where trustors choose to lower their trust following a trust violation; and trust restoration, where trust stops diminishing after a disruption and may eventually be restored. In the early stages of a relationship, trust in a system is primarily based on the system's performance and perceived reliability. Trust can shift in response to changes in the system's performance (Lee & Moray, 1992; Muir, 1994). These variations in trust are often positively correlated with changes in automation use. As trust diminishes, operators may resort to more frequent manual control and consequently, less inclined to test out the automation's capabilities. When operators continue to interact with the system, they start forming expectations of future performance of the automation. This "expectation" can correspond to "anticipation" in Sheridan's (2019) model. In addition, extended interaction leads to the formation of generalizations about the automation performance and broader positive beliefs about the system's behavior, referred to as "faith".

Trust at the current moment is influenced by the trust at the previous moment (Lee Moray 1992) in other words, trust is history-dependent where the initial level of trust is determined by past experiences in alike situations (Lee & See 2004). The anchoring effect, which describes how individuals make estimates or decisions starting from an initial reference point, can significantly impact trust dynamics. Lewicki & Brinsfield (2011) indicated that an operator's initial trust level acts as an anchor, substantially influencing their subsequent trust assessments and reactions to automation performance, particularly after system malfunctions (Lee & See, 2004; Merritt & Ilgen, 2008). The degree to which trust is eroded following a malfunction is often proportional to this anchored initial trust level.

The initial trust and trust expectations (anticipation) are not static but rather influenced by an individual's prior experience or inherent propensity to trust, which varies from person to person. As a result, individuals who commence their interaction with an automated system with high trust expectations are often more sensitive to any changes in the system's performance (Pop et al., 2015). A significant system malfunction can dramatically alter their trust levels, resulting in a steep decline compared to those who started with lower trust expectations.

A. Model structure

Figure 2 captures the trust dynamics with three key stocks, *Trust*, *Perceived Performance*, and *Expectation of Performance*. The focal point of this model is the *Trust* stock. Adopting Caddell and Nilchiani (2023) formulation of interpersonal trust, trust accumulates changes based on the difference between the *Expectation of Performance* and the *Perceived Performance*. Let $T(t)$ be the *Trust* stock at time t and $c(t)$ be the *Change in TiA*. Using this convention, Equation (1) allows us to articulate the formulation of the stock. *Change in TiA*, *C(t)*, is articulated in Equation (2), where ρ denotes the *Initial Trust*, θ denotes the *Expectation Gap*, ω indicates the *Difference Between the Maximum TiA (Faith)* and *Current Trust*, and λ represents the *Trust Adjustment Time*.

$$T(t) = \rho + \int c(t)dt \quad (1)$$

$$c(t) = \min\left\{\frac{\theta}{\lambda}, \frac{\omega}{\lambda}\right\} \quad (2)$$

Based on this formulation, trust at time t is equivalent to the initial level of trust and an accumulation of all the changes in trust during the interaction period. This formulation sets the basis of the model.

The model posits that an individual's $Initial\ Trust$ (ρ) is influenced by their previous experiences in similar contexts and their inherent disposition or propensity to trust, as outlined by Merritt and Ilgen (2008) and considered a source of intra-individual variability. While individual characteristics significantly impact trust levels, these traits are generally stable and relatively constant over time. In our model, individual characteristics primarily influence two aspects: (1) Initial Levels of Trust (Stocks), or an individual's baseline trust, is

shaped by their trust propensity. This means different individuals may begin interactions with varying levels of trust and expectations based on their unique predispositions and prior experiences. (2) Adjustment Times, where individuals differ in their response rates to cues that might influence trust. Some might adjust their trust levels quickly in response to new information or experiences, while others may do so more slowly. These variations affect how rapidly or gradually individuals' trust levels change in response to perceived performance. Despite these individual differences in initial trust and adjustment rates, the underlying feedback structure that governs trusting behavior remains consistent across individuals.

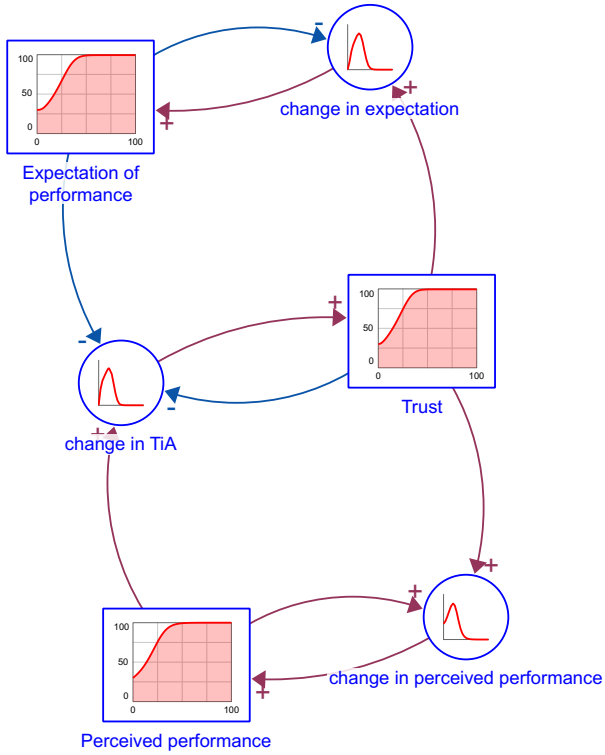


Figure 2. Simplified Causal Loop Diagram (CLD) of the model.

The model consists of three prominent feedback loops responsible for trust patterns. The reinforcing feedback loop (R1), as shown in Figure 3, is initiated by an elevation in the trust, heightening the likelihood of generating *Trusting Behavior* and use of the system. While increased trust does not always translate into trusting behavior, *Desire to Allocate Trusting Behavior* represents a partial contribution in this regard. *Desire to Allocate Trusting Behavior* can be influenced by the *Perceived Risk* of using the autonomous system. In scenarios where negative outcomes are more probable, there is an increased reluctance among individuals to engage with automated systems. Furthermore, following a decline in trust, the duration required for individuals to restore their trust is more extended in high-risk circumstances than in those with lower risks (Riley, 1994). That said, when a system performs well, there is an overall tendency to allocate more tasks. This, in turn, results in testing the *System's Performance*, informing the *Perception of Performanc*e and consequently updating an individual's trust.

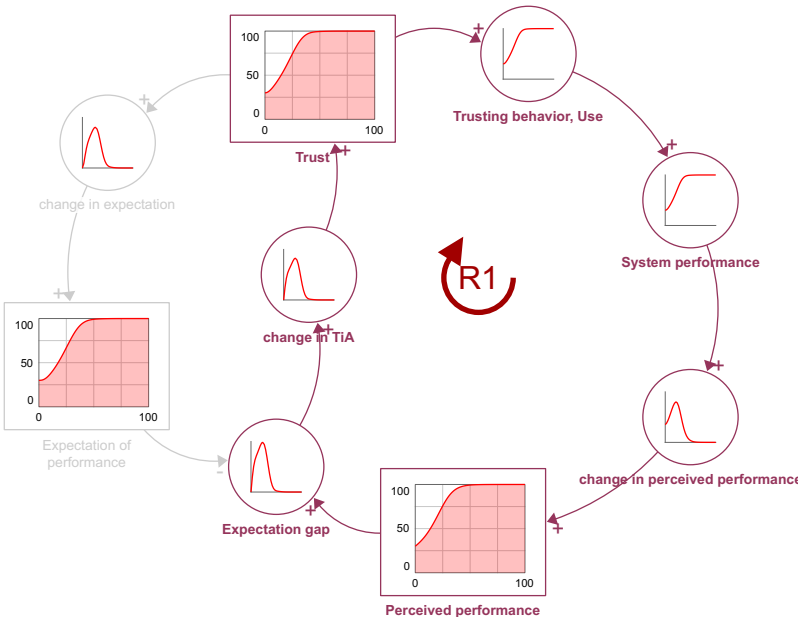


Figure 3. Trust formation reinforcing feedback loop

The balancing loop (B1) starts by forming the *Expectation of Performance*, and subsequently impacts *Trust* through the discrepancy between the *Perceived Performance* and *Expectation of Performance*, labeled as *Expectation Gap* in Figure 4. This loop allows trust to adjust to the observation of the system's performance. An increase in trust may raise the user's expectations. If the perceived performance does not meet these heightened expectations, a gap between expectation and perceived performance occurs. The widening gap can erode trust, creating a negative feedback loop.

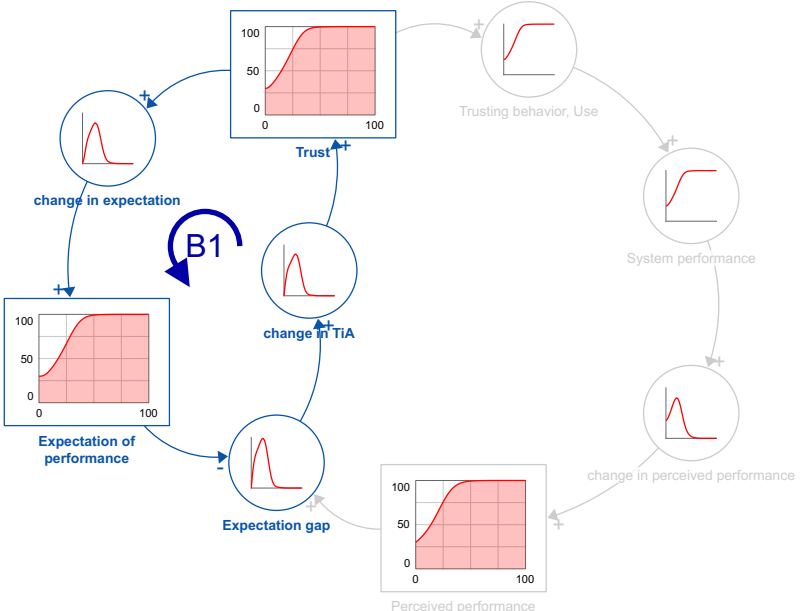


Figure 4. Expectation balancing feedback loop.

The second balancing loop (B2) controls the level of trust and highlights *Limits to Trust*. With the increase in trust, the gap between the current level of trust and its maximum level

(*Faith*) diminishes, leading to a reduced variation in trust and constraining its further growth. This loop establishes a first-order control mechanism that guarantees trust cannot surpass the maximum limit; there is a cap on how much one can trust a system.

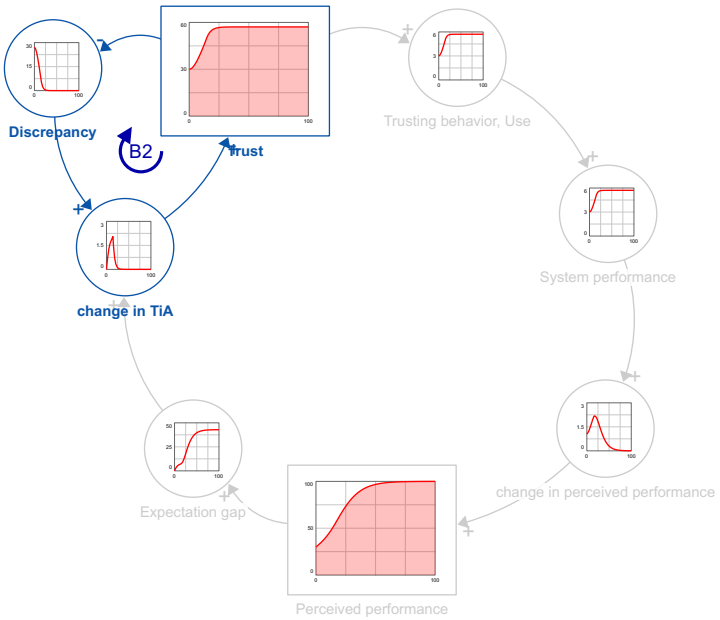


Figure 5. Limits to trust balancing feedback loop.

System Malfunction in this model represents a sudden disruption in the *System Performance*, formulated with a Pulse function that includes the *Magnitude* and *Time* of the *System Malfunction*. On the other hand, *System Capability* reflects the limiting cap to the *Perceived Performance*. This simply means what automation can or cannot do in different situations. *Environmental Challenges*, such as adverse weather conditions or limited visibility, can greatly impact the *System Capability*.

3.4 Model simulation

The simulation exercises were conducted utilizing the STELLA Architect version 3.5 software, spanning 100 time steps, with a fractional delta time (DT) set at 1/125, and employing the Euler method for integration. The outcomes of these simulation runs are analyzed and discussed in the following sections, under diverse scenarios.

A. Baseline equilibrium (perfect automation)

At equilibrium, when the system maintains some *Nominal Capability* level that aligns with the operator's *Initial Trust* (e.g., 30 on a 100 scale), there is no observable change; the model remains in a state of equilibrium (See Figure 6). However, altering the initial values or the *Nominal Capability* yields significantly varied behaviors. Deviation from the equilibrium activates the reinforcing loop, leading to goal-seeking dynamics, as illustrated in Figure 7. According to these dynamics, operators identified with higher trust levels exhibit more trust behaviors to test out the system's performance and accumulate more positive *Perceived Performance* (given perfect automation). Conversely, those with trust below the nominal level trigger these reinforcing loops negatively, leading to path dependency, a phenomenon

where the structure of the system and its initial condition drive its trajectory (Sterman 2002). Such path-dependent dynamics in trust evolution have been noted in various studies (Castelfranchi & Falcone, 2010; J. D. Lewis & Weigert, 2012, 2012), emphasizing that the past and initial states of trust significantly shape future trust levels and behaviors. The manifestation of path dependence in trust dynamics is further influenced by the dominance of specific loops at given times. As conditions change, different feedback loops may become more influential, shaping the direction and rate of trust evolution. This means that the system's behavior is not only determined by its initial state but also by how different reinforcing or balancing loops interact and dominate over time, leading to complex and sometimes unexpected paths of trust development. Figure 8 illustrates the relative dominance of various feedback loops over time, depicted as a percentage of the total effect. The dashed horizontal line at 50% represents a threshold or reference point, indicating a balanced influence where neither the reinforcing nor the balancing loops are dominant. The influence of different feedback loops varies over time, with periods where the system seeks balance (as indicated by the B loops) and other times where it experiences growth or decline (as indicated by the R loop). The Bu1- and Bu1+ loops suggest that the same balancing loop might have different phases or actions depending on other conditions in the system.

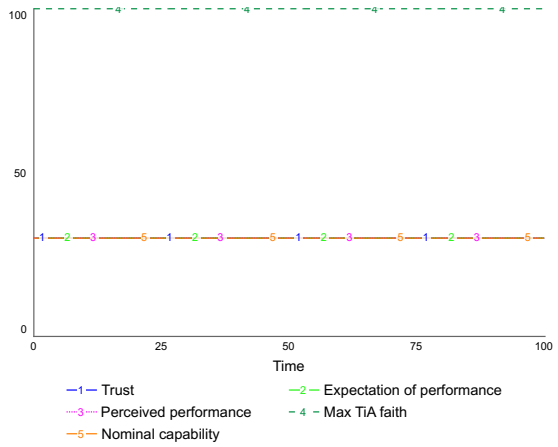


Figure 6. Baseline equilibrium of the model.

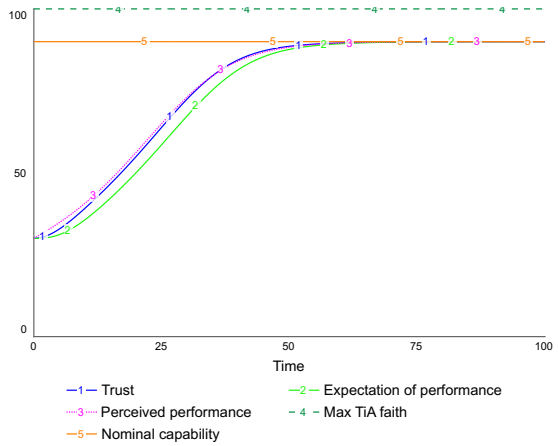


Figure 7. S-shape goal-seeking behavior.

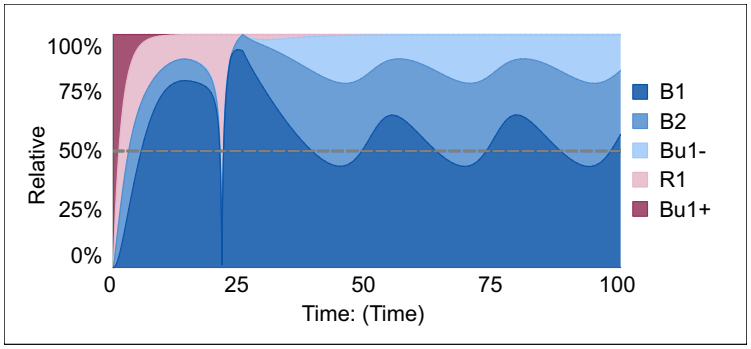


Figure 8. Relative dominance of feedback loops; B1: Expectation gap, B2: Change in Expectation, R1: Trust formation, Bu1∓: Perceived performance

The loop dominance and path dependence behavior in the model resonates with everyday experience, as higher levels of trust tend to demonstrate more reliance on the system and, if capable, build upon that trust positively to a certain limit. In contrast, incompetent automation faces diminishing trust behaviors and ultimately a reduction in trust-based interactions. Although formal validation of these dynamic behaviors requires historical data as reference mode, the model confirms that the structure of the model which is based on the existing literature can produce the behavior described and suggested in the literature.

The model also responds to variations for different inputs. Figure 9 illustrates the sensitivity analysis for random and normally distributed inputs for the *Nominal Capability*. As anticipated, the path dependence appears at both extremes.

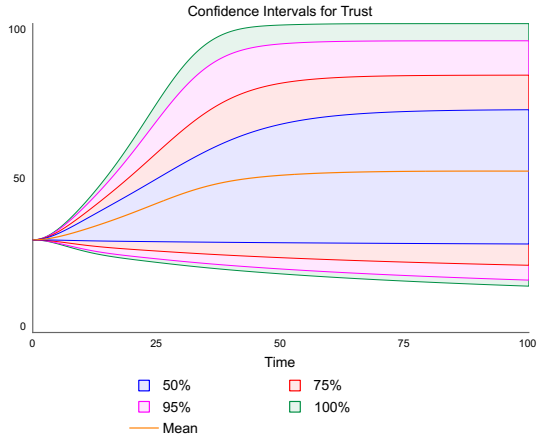
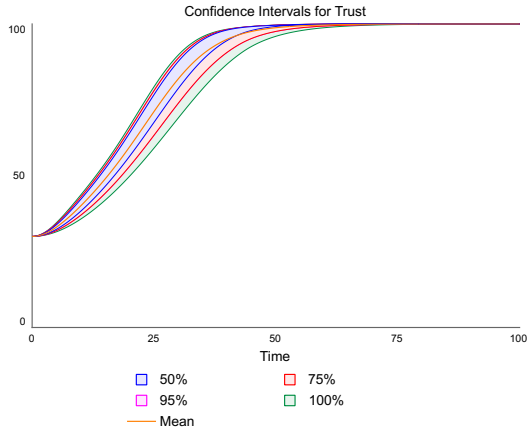


Figure 9. Sensitivity analysis for various inputs of nominal capability

B. Individual variability

The baseline simulation result has demonstrated a path dependency in individuals' trust development trajectories. As outlined earlier, variations among individuals manifest in the form of *Initial Trust* values and the temporal duration required for adjustments. Essentially, the model's gradient elucidates the rate at which trust is established or eroded, reflecting the velocity of trust accumulation or decay. These intra-individual dynamics are depicted in Figure 10, illustrating that while distinct levels of *Propensity to Trust* lead to diverse outcomes, they nevertheless follow comparable patterns of trust development. Likewise, the model exhibits convergence toward its terminal state of path dependency, influenced by

varying *Initial Trust* values, as demonstrated in Figure 11. This convergence is attributed to adjustments arising from the discrepancy between expected and perceived performance. When perceived performance meets or exceeds expectations, trust incrementally escalates.



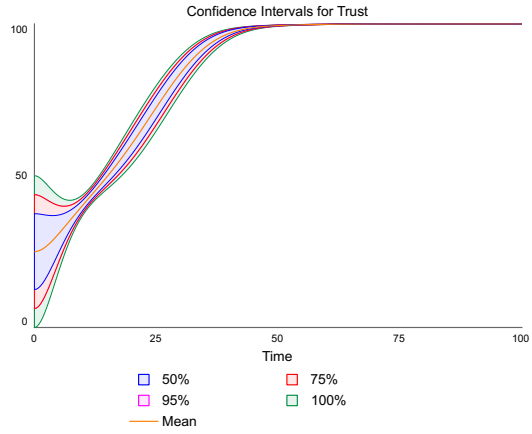


Figure 10. Sensitivity analysis for various inputs for propensity to trust
Figure 11. Sensitivity analysis for various inputs for initial trust

C. Initial values mismatch (Higher expectations)

Variability in outcomes is also influenced by the disparity in initial values between *Expected* and *Perceived Performance*. When an individual's anticipated performance is significantly higher than what is observed, this discrepancy activates a negative feedback loop, precipitating an initial decline in trust in automation. Consequently, with diminished trust, operators are slower to engage with the automation system and gradually perceive its effectiveness, leading to a protracted period of trust accumulation and expectation adjustment. As the gap between expectation and perception narrows, trust begins to accumulate, albeit at a diminished velocity, as shown in Figure 12.

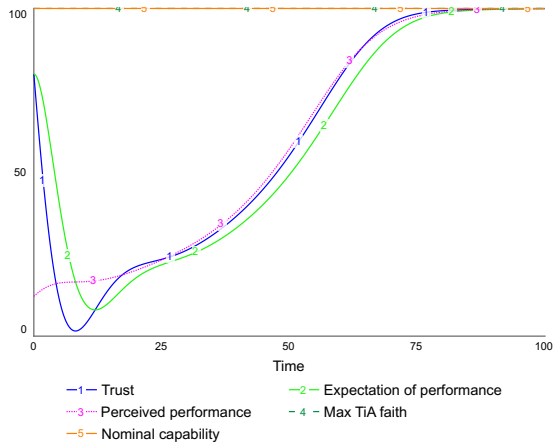


Figure 12. Mismatch in initial values of perceived performance, trust, and expectation of performance

D. Perceived risk

As previously noted, a critical constraint in automation use is the *Perceived Risk*. Specifically, some scenarios prompt operators to favor reduced use of automation due to

associated risks. The impact of this risk-averse behavior is depicted in Figure 13, where varying levels of *Perceived Risk* alter the trust development trajectory. Such high-risk situations restrict engagement with the autonomous system, impeding the observation and assessment of its performance and thus leading to a more slow and gradual accumulation of trust.

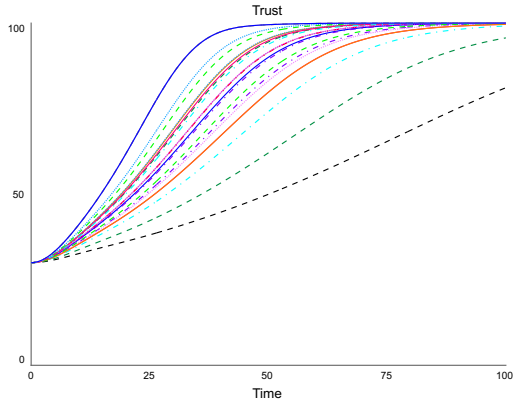


Figure 13. Multiple simulation runs with varying inputs for perceived risk

C. System malfunction (imperfect automation)

So far, simulations have assumed optimal system performance, leading to a natural increase in human trust through sustained usage. However, to understand trust dynamics in response to system malfunction, we introduced a disruption into the baseline model using a Pulse function. This function simulates an error with a 50% *Magnitude* occurring at time $t = 25$, causing an abrupt decline in both *System Performance* and *Perceived Performance*, as illustrated in Figure 14. The resultant breach in expectations causes a delayed but significant decrease in trust, maintaining it at a lower level until system performance begins to recuperate and align with the operator's expectations.

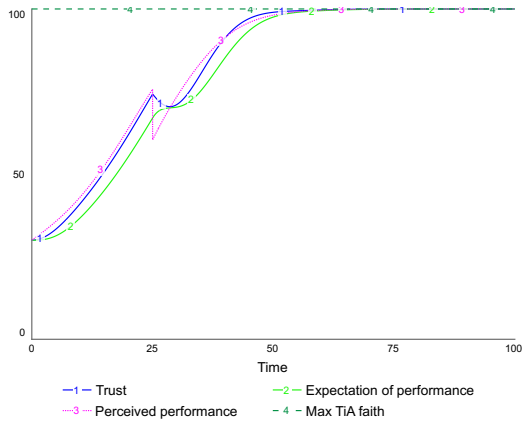


Figure 14. Perceived performance and trust behavior after a system malfunction at t=25

The influence of system malfunction on trust is not immediate, but rather a gradual decline in trust over time. Similarly, the recovery process after malfunctions is not instantaneous but unfolds over a period (Yang et al., 2017). This can be observed in Figure 14 where trust experiences a lower recovery level than the upward perceived performance.

The timing of system malfunction plays a pivotal role in the process of trust development. In a subsequent simulation, we applied the same *Error Magnitude* (50%) but introduced it earlier at $t = 12$. This adjustment resulted in a more pronounced drop in trust followed by an extended recovery period (Figure 15). A plausible explanation for this dynamic is that breaches of trust during early interactions cause operators to become more hesitant to engage with the system, thereby delaying their ability to observe its potential performance and correspondingly prolonging the duration required to rebuild trust.

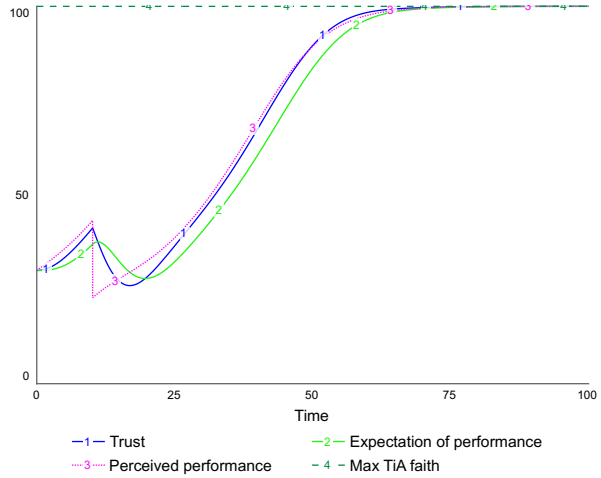


Figure 15. Perceived performance and trust behavior after a system malfunction at t=12

We simulated the model with several different timings of the *System Malfunction*. As Figure 16 illustrates the timing of errors significantly influences trust recovery in an automated system. The topmost line, which does not exhibit any drop, represents the baseline scenario without any system errors. Early errors, introduced at around time $t = 10$, result in a steep decline in trust and a prolonged recovery phase where trust does not reach the level of the no-error scenario within the most observed time frame. Errors occurring mid-way, around time $t = 25$, still cause a notable dip in trust but allow for a quicker rebound than early errors. In contrast, errors introduced later in the process have a relatively minor impact on trust levels, with a faster return to near-baseline trust.

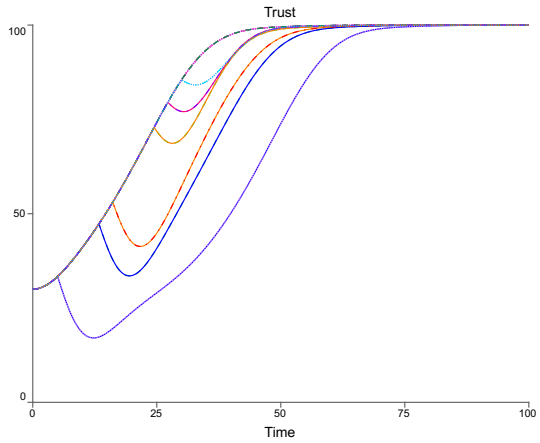


Figure 16. Multiple simulation runs with varying inputs for error time

In our investigation into the impact of system malfunctions on trust, we conducted an empirical study. This approach was motivated by two primary factors. Firstly, the temporal

dimensions of trust dynamics—encompassing the formation, erosion, and restoration of trust—necessitate meticulous experimental design with different timeframes. Given that current literature does not provide definitive timelines for the establishment of trust or recovery post-malfunction, it was sensible to conduct studies that simulate rapid-onset stimuli. Such conditions are conducive to observing trust changes over shorter durations, thereby facilitating a more accurate assessment of malfunctions' effects on trust. Secondly, it is pragmatic to recognize that automated systems are not infallible and that malfunctions, while undesirable, are a realistic aspect of their operation. Therefore, integrating these realistic elements into our experimental design offers valuable insights into trust dynamics within authentic operational contexts.

4 Experiment

The objective of this study is twofold. *First*, to empirically test a segment of the proposed System Dynamics model specifically tailored for application in Maritime Autonomous Surface Ships (MASS). *Second*, to investigate the behavioral manifestations of trust in the context of monitoring automation, utilizing eye movement data. The focus is on assessing the frequency of monitoring and analyzing fixation metrics to understand how system malfunctions influence the operator's engagement with the automated system.

A. Participants

A total of 30 participants, consisting of nautical students and licensed instructors from a Norwegian maritime university were recruited for this study. Participant recruitment was accomplished via non-probabilistic snowball sampling (Vehovar et al., 2016). The 22 male and 8 female participants were between the ages of 18 and 55 years. Table 1 provides an overview of 30 participants recruited in this study.

Prior to their participation, all participants provided written informed consent, acknowledging their right to withdraw from the study at any point. Upon the conclusion of their involvement, participants were compensated with a gift card. The experimental procedures, including the management of data, received approval from the Norwegian Centre for Research Data (NSD) under project number 407324. In compliance with NSD guidelines, all personal data acquired during the study was handled with stringent confidentiality and security measures.

TABLE 1: Participants' Demographic information

		n	*%*
Age	18-24	16	53.3
	25-34	11	36.7
	35-44	1	3.3
	45-54	2	6.7
Gender			
	Female	8	26.7
	Male	22	73.3
Education			
	High School	5	16.7
	Bachelor	17	56.7
	Master	6	20.0
	Doctorate	2	6.7
Seafaring Experience			
	None	16	53.3
	1-5	7	23.3
	5-10	4	13.3
	10-15	2	6.7

	Over 15	1	3.3
Navigation knowledge	Very poor	1	3.3
	Some	6	20.0
	Good	15	50.0
	Very good	6	20.0
	Expert	2	6.7

Note. N = 30

B. Apparatus

The Kongsberg K-Sim desktop bridge training simulator was used as a test environment for the design of scenarios and simulation of an autonomous vessel. The simulator provides maritime navigators with a platform to refine their navigational skills through practice and training. K-Sim supports the creation of maritime traffic scenarios, which can be simulated in real time, offering an immersive and dynamic training environment (Kongsberg, 2023). As shown in Figure 17, the experimental setup consists of three screens, a radar information screen, a bridge (main interface), and an Electronic Chart Display and Information System (ECDIS).

C. Data collection

Quantitative data collection was performed through self-reported questionnaires and metrics from the eye-tracking glasses. Demographic information of the participant was collected through a customized questionnaire. Personality traits data were collected using a 50-item International Personality Item Pool (IPIP) for the Big-Five personality factors (Goldberg et al., 2006). Trust in Automation (TiA) and perceived reliability data were gather via trust questionnaire developed by Körber (2019). Eye movement data were collected using Tobii Pro Glasses 2 and the Tobii Pro Lab software. The sampling rate of the eye-tracking glasses was 50 Hz.

The raw data collected from eye-tracking were utilized to compute several metrics, which are categorized into three established groups (Boudreau et al., 2009; Lu & Sarter, 2019; Yang et al., 2017): temporal, spatial, and count metrics. Temporal metrics include total and average fixation duration, with a fixation characterized as a relatively stationary gaze within a certain dispersion threshold (around 2°), lasting for a minimum duration (usually between 100-200 ms), and a velocity below a set threshold (commonly 15-100°/s) (Jacob & Karn, 2003). The spatial metrics, which consist of mean saccade amplitude, backtrack rate, rate of transitions, and scan-path length per second, address the efficiency and pattern variability of eye movements. These metrics serve to quantify the eye's movements and the strategy of visual information intake. The count metrics are concerned with the number and frequency of fixations and transitions between predefined Areas of Interest (AOI). In this context, an AOI is a specified zone determined by the experimenter to focus the analysis of eye movement data. For this experiment, two AOI sets were employed: one set considered the alarm section on the bridge monitor and the mid-section on the ECDIS collectively to investigate the impacts of system malfunction on the eye movement metrics; the second set included the three screens as AOI to visualize the participants' scanning behaviors.

The eye movement data was primarily processed using the Tobii Pro software package, IBM SPSS Statistics software (version 29), and R programming language.

Figure 17. Experimental setup: Radar, Bridge (main interface), ECDIS

D. Design

This study implemented an observational within-subject design where all participants were exposed to the same stimuli. A mid-size Ro-Ro vessel was selected, sailing autonomously from the port of Horten to Moss in Norway. Three additional vessels were added to the simulated environment to emulate a realistic traffic situation with one vessel crossing situation, see Figure 18. The vessel was initiated in an autonomous navigation mode (N) set to 100% speed, following a pre-validated course marked by specific waypoints. Participants were tasked with closely observing the vessel's navigational performance, which included route monitoring, navigation control, and collision avoidance, to ensure its safe passage. They were provided with the option to intervene and control the vessel by switching to manual control (M) or autopilot (A) mode, as well as by modifying the vessel's speed if deemed necessary. In the event of any system malfunctions, participants were expected to respond to alarms and execute corrective maneuvers to return the vessel to its designated trajectory. The duration of the study was approximately 40 minutes.

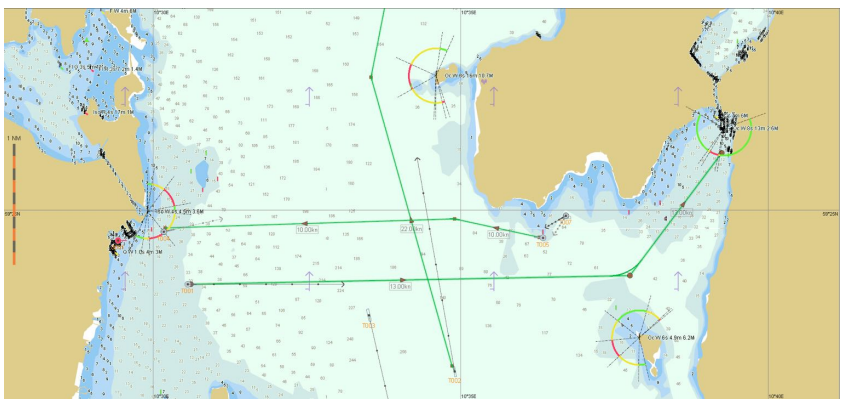


Figure 18. Simulated vessel traffic environment

E. Procedure

Figure 19 delineates the experimental sequence. At the outset, a brief introduction to the goal of the study was provided. Following this introduction, they proceeded to read and sign the informed consent form. The demographic and personality questionnaires were administered, and the eye-tracking device was calibrated to ensure accurate data collection tailored to each participant. Participants were then familiarized with the interface and setup and trained on the tasks they would be required to perform during the session.

In the initial phase of the study, the system operated flawlessly, exemplifying perfect automation by maintaining accurate control and adherence to the predetermined route. During the latter half of the session, the system was programmed to introduce a malfunction by halting one of the steering gear pumps. This malfunction led to the vessel's deviation from its charted course, subsequently activating an alarm to notify the participant of the issue. Deviation from the course commenced after 10 minutes and lasted for 60 seconds. If the test subject failed to notice and/or take over, the vessel was set to go back on course after the error period. Test subjects were expected to notice the change in course anytime within the 60 seconds. By the end of the study, if any test subject did not notice the deviation, it would be considered a failure to recognize. The self-report trust measures were collected after 10 minutes of interaction with the system (pre-error) and at the end of the session (post-error).

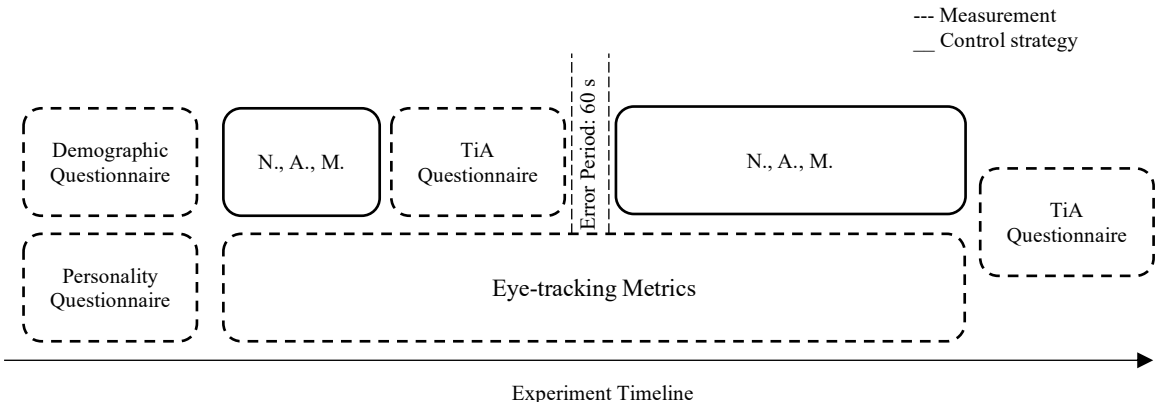


Figure 19. Sequence of experiment

F. Analysis and results

Data analysis embarks by employing visualization techniques to enhance the understanding and presentation of the collected data, particularly to depict the changes of trust in automation between two distinct points in time, as shown in Figure 20. During the data collection process, it was observed that two participants (participants ID. 21 and 28) did not register the introduced error, leading to an anomalous increase in their trust in the automated system. Consequently, these two participants were omitted from the subsequent data analysis, as their experiences did not accurately reflect exposure to the critical stimulus — the system malfunction.

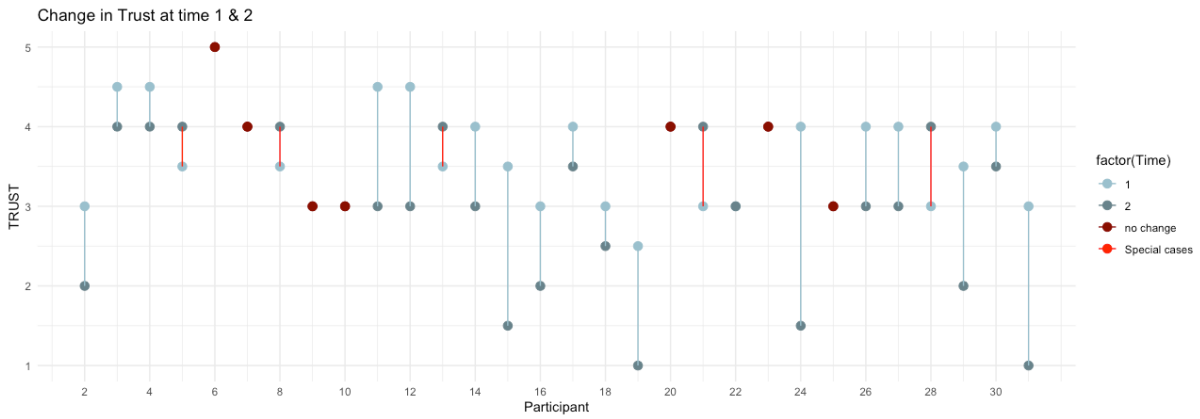


Figure 20. Trust in Automation at two time points for 30 participants

Figure 21 provides heatmaps representing the visual attention allocation of participants across three main displays: Radar information, Bridge Display System (middle), and Electronic Chart Display and Information System (ECDIS). In the pre-error period, visual attention is distributed across all three displays, with a relatively even distribution of fixation duration indicating routine monitoring behavior. However, in the post-error period, there is a marked shift in visual focus towards the Bridge Display and ECDIS. The increased intensity and concentration of colors on these displays in the heatmap signal a heightened attentional demand, likely due to participants seeking critical navigational information. This shift suggests a cognitive reallocation of resources towards areas perceived as more relevant for the post-error period, reflecting an adaptive response to operational anomalies.

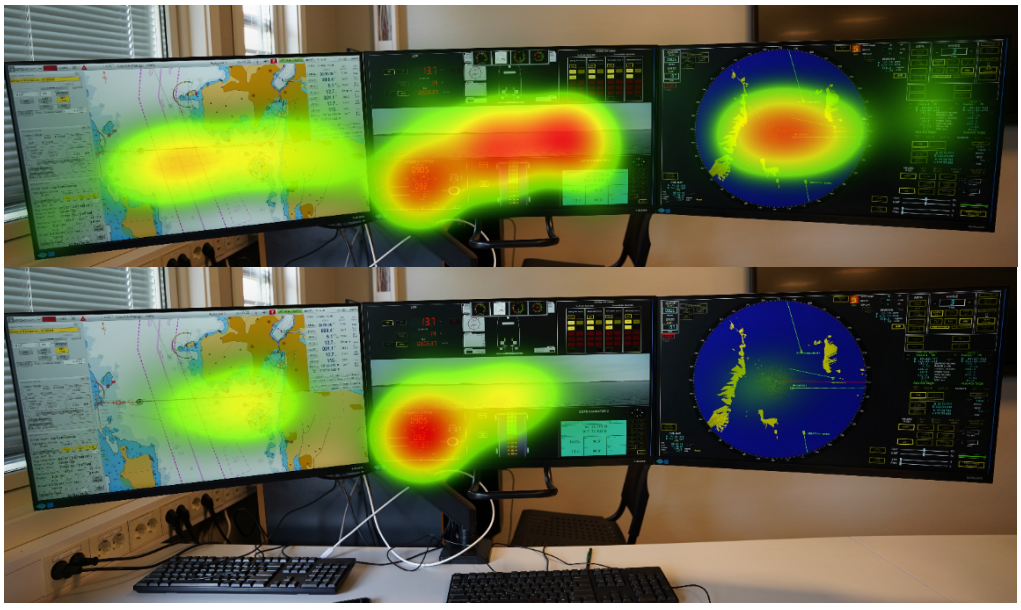


Figure 21. Heatmap of visual attention before and after the system malfunction

A paired samples t-test was employed to assess the differences in various gaze metrics as well as perceived reliability and trust in automation before and after the error occurrence. As illustrated in Table 2, temporal metrics displayed substantial increases in the total duration of fixation (TDF) post-error (M_{diff} = -56082.3, 95% CI [-77614, -34550], Cohen's d = 1.01) and

average duration of fixation (ADF) (M_{diff} = -68.6, 95% CI [-110, -26], Cohen's d = 0.63), suggesting an extended focus on AOIs after the error. Spatial metrics did not show a significant change in the average amplitude of saccades (AMS), but there was a notable increase in the total amplitude of saccades (TAS) (M_{diff} = -972.5, 95% CI [-1463, -481], Cohen's d = 0.77). For count metrics, there were significant increases in the number of visits (NV) (M_{diff} = -56.4, 95% CI [-70, -42], Cohen's d = 1.53), number of saccades (NS) (M_{diff} = -35.1, 95% CI [-53, -16], Cohen's d = 0.73), and number of fixations (NF) (M_{diff} = -119.6, 95% CI [-162, -76], Cohen's d = 1.07). These results indicate that following the system malfunction, participants engaged with the information from the AOIs more intensively. Furthermore, perceived reliability (REL) and trust in automation (TRU) significantly decreased from pre-error to post-error, with REL showing a larger mean decrease (M_{diff} = 0.77, 95% CI [0.53, 1.01], Cohen's d = 1.25) compared to TRU (M_{diff} = 0.66, 95% CI [0.34, 0.98], Cohen's d = 0.62).

<Table 2 Here: Check the last page>

Next, to delineate the relationships between perceived reliability, trust in automation, and various personality traits across two time points, a correlation analysis was performed. As illustrated in Table 3, perceived reliability (REL) and trust in automation (TRU) at time 1 (t_1) have a strong positive correlation (r = .65, 95% CI [.37, .82]), which suggests a robust association that strengthens at time 2 (t_2), as indicated by an even higher correlation (r = .79, 95% CI [.59, .90]). Change in trust in automation (dTRU) and the change in perceived reliability (dREL) are also positively correlated (r = .62, 95% CI [.33, .81]), indicating a concurrent decrease in trust and perceived reliability from the two time points. It is important to note that the Big Five personality traits do not show significant correlations with changes in trust, suggesting that personality may not play a substantial role in the observed decline in trust because of system malfunction. The internal consistency of our scales, as measured by Cronbach's alpha, ranges from acceptable (.69 for Conscientiousness) to excellent (.93 for TRU at t2), ensuring the reliability of our measures.

<Table 3 Here: Check the last page>

To align the empirical findings with the segment of the System Dynamics model under investigation, a comparative analysis was conducted. This involved measuring the percentage changes in empirical trust in automation and perceived reliability against the decline in perceived performance and trust as depicted in the model. By quantifying these changes, we aimed to assess the model's accuracy in reflecting real-world trust dynamics within automated systems. The empirical results showed that perceived reliability (REL) experienced a mean decrease from 3.46 to 2.69, constituting a 15.4% decrease when normalized against the maximum Likert scale value of five, (0.77 / 5) * 100 = 15.4%. Similarly, trust in automation (TRU) had a mean reduction from 3.68 to 3.02, translating to a 13.2% decrease, (0.66 / 5) * 100 = 13.2%. These empirical changes in trust and perceived reliability were put into perspective by comparing them with the simulation model's predictions. The model indicated a 14% (from 46.58 to 32.50) decline in perceived

performance for the period of 12-17 and a parallel 12.7% (from 44.7 to 32) decrease in trust in automation. This comparative analysis demonstrates a congruence between the simulated and observed trust dynamics, validating the model's ability to reasonably reflect the impact of system malfunctions on trust and reliability. Despite variations in timing and magnitude, the alignment of trends confirms the model's relevance in simulating trust behavior in the wake of system malfunctions, albeit with a recognition of its limitations and scope for refinement.

5 Discussion

The results from the system dynamics simulated scenarios highlight the critical role of initial conditions in shaping long-term trust dynamics in human-automation interactions, corroborating the research on the lasting impact of first impressions. Importantly, these outcomes underline the necessity of aligning initial conditions for system performance, user trust, and performance expectations. This alignment is pivotal for the effective operation of autonomous systems. The idea that building trust extends beyond the mere enhancement of system performance to include the alignment of user expectations with the system's realistic capabilities has reflective design implications. Successful automation systems must not only be technically competent but also effectively communicate their capabilities to users to set realistic expectations. It is also essential for designers to align system capabilities with the tasks they are intended to perform. Overestimating these capabilities can lead to scenarios where the system fails to meet expectations, triggering a negative feedback loop of diminishing trust. One approach would be through transparency about system capabilities, and providing users with clear, accurate information about what the system can and cannot do. User training should also emphasize realistic expectations about the system's capabilities and limitations, helping to prevent the erosion of trust due to misunderstandings or unrealistic expectations.

The phenomena of loop dominance and path dependence, as depicted in the model, demonstrate that when a system is trusted and proves itself capable, trust tends to grow, leading to increased reliance on the system. This positive reinforcement loop suggests that trust, once established, can be self-perpetuating up to a certain threshold, provided the system continues to meet or exceed expectations. Conversely, when automation is perceived as incompetent or unreliable, the model shows a negative feedback loop. In these cases, trust begins to diminish, and as a result, trust-based interactions with the system decrease. This decline in trust can lead to a reduced willingness to rely on the system, with users potentially reverting to manual controls or alternative methods. The diminishing trust is a critical concern, as it not only affects current interactions but can also have long-term implications for the acceptance and usability of automated systems.

The study's exploration of trust recovery dynamics in the aftermath of system malfunctions provides valuable insights into the resilience of trust in automation. A key finding is the important role that the timing of a malfunction plays in shaping trust trajectories. Malfunctions occurring early in the interaction with an automated system result in a more pronounced decrease in trust, accompanied by a prolonged recovery period. This observation underscores the crucial need for establishing and maintaining reliability from the onset of

system usage. These findings align with the concept of the anchoring effect in trust relationships, where the initial level of trust sets a baseline that significantly influences future trust dynamics. In the context of human-automation interaction, early experiences with the system serve as an anchor point. If these initial experiences are negative, such as encountering early malfunctions, they set a lower trust baseline, making it more challenging to build trust subsequently. Conversely, positive early experiences can establish a higher baseline, making the system more resilient to future setbacks. For designers and operators of automated systems, this emphasizes the importance of robust initial system testing and quality assurance to minimize early malfunctions. Ensuring that the system operates reliably when first introduced to users can create a strong foundation of trust, which is more likely to withstand future issues. Additionally, users should be informed about the potential for malfunctions and the procedures for addressing them. This transparency can mitigate the negative impact on trust when malfunctions do occur.

The empirical findings from the study on Maritime Autonomous Surface Ships (MASS) align with the established research body, including the works of Lee and Moray (1992), Moray et al. (2000), and Yang et al. (2016, 2017), reinforcing the principle that trust in automation is closely tied to the system's performance outcomes. Consistent with these prior studies, our research observed a decline in trust following a system malfunction. Notably, the study reveals that personality traits do not directly influence perceived reliability or variations in trust. This finding corroborates the simulation model's hypothesis that individual differences primarily affect initial trust levels and adjustment times rather than ongoing trust perceptions.

6 Conclusion and future work

This study developed a system dynamics model of trust in automation to demonstrate the utility of simulation modeling for creating dynamic, testable hypotheses about trust and trusting behaviors. Unlike static conceptual models, the proposed model provides a flexible tool that can be adjusted and validated against empirical data. This flexibility allows researchers to explore a range of conjectures about how trust forms and evolves over time. The adaptability of the model is particularly beneficial for future research endeavors. It provides a robust tool that can be tailored to investigate trust in a variety of contexts by manipulating values or refining the structure of the model. This adaptability is crucial for modeling complex human-automation interactions across diverse operational environments. Furthermore, employing the system dynamics approach enabled the construction of a model based on dynamic feedback loops. This approach is instrumental in replicating the non-linear and often reciprocal nature of trust behaviors. By considering the interplay between reinforcing and balancing loops, the model can generate behaviors that more accurately reflect the fluctuations and transitions observed in real-world trust scenarios.

The proposed model primarily focuses on the structural aspects of trust dynamics, rather than specific contextual details. This approach brings both strengths and limitations. On the one hand, the absence of explicit contextual elements restricts the model's direct applicability to specific scenarios. Each unique situation, with its distinct variables and conditions, may require adjustments to the model's numeric sensitivities for accurate representation and

prediction. On the other hand, the model's general structural design and its depiction of path-dependent behavior are broadly applicable across a range of scenarios. This generalizability is one of the model's key strengths. The underlying structure of the trust relationship – characterized by its feedback loops and the overall dynamics of trust formation, dissolution, and restoration – is likely to remain relevant in various contexts.

For system designers, this model can serve as a guide for creating systems that not only meet technical and functional requirements but also address the psychological and behavioral aspects of trust. It underscores the importance of considering user perceptions, experiences, and responses to automated systems from the initial stages of design. This consideration is central to ensuring that users not only trust the system's capabilities but also understand its limitations, thereby fostering a balanced and informed trust relationship.

It is important to note that different Levels of Automation (LOAs) can significantly influence the trust dynamics (Walliser, 2011), especially during the monitoring process, the focus in this paper was narrowed to scenarios involving the highest level of automation, such as Full Self-Driving (FSD) in vehicles or Auto-track mode in maritime vessel navigation systems, where the operator's primary role is relegated to supervising the automation system and intervening when necessary. While the model is adaptable to any given LOA, it does not explicitly account for transitions between different LOAs or the intricate relationships between trust levels across various automation levels. This limitation was a deliberate scope constraint to maintain focus and clarity but does suggest an area for potential future expansion and refinement of the model.

Future iterations of the model could also incorporate modular elements that allow for easy adaptation to different contexts. This could include features that adjust for varying levels of system reliability, user experience, environmental factors, or other relevant contextual variables. Such enhancements would make the model more versatile and applicable to a broader range of specific situations, thereby increasing its usefulness in practical applications.

Future research may focus on applying the model to specific contexts, empirically testing its assumptions, and exploring how variations in individual characteristics may influence trust adjustment times.

Declaration of Competing Interest

The authors report there are no competing interests to declare.

7 References

Akash, K., Wan-Lin Hu, Reid, T., & Jain, N. (2017). Dynamic modeling of trust in human-machine interactions. *2017 American Control Conference (ACC)*, 1542–1548. https://doi.org/10.23919/ACC.2017.7963172

Azar, A. T. (2012). System dynamics as a useful technique for complex systems. *International Journal of Industrial and Systems Engineering*, *10*(4), 377. https://doi.org/10.1504/IJISE.2012.046298

Barber, B. (1983). *The logic and limits of trust*. Rutgers University Press.

Basu, C., & Singhal, M. (2016). Trust dynamics in human autonomous vehicle interaction: A review of trust models. *2016 AAAI Spring Symposium Series*.

Boubin, J. G., Rusnock, C. F., & Bindewald, J. M. (2017). Quantifying Compliance and Reliance Trust Behaviors to Influence Trust in Human-Automation Teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 750–754. https://doi.org/10.1177/1541931213601672

Boudreau, C., McCubbins, M. D., & Coulson, S. (2009). Knowing when to trust others: An ERP study of decision making after receiving information from unknown people. *Social Cognitive and Affective Neuroscience*, *4*(1), 23–34.

Caddell, J. D., & Nilchiani, R. (2023). The Dynamics of Trust: Path Dependence in Interpersonal Trust. *IEEE Engineering Management Review*, *51*(3), 148–165. https://doi.org/10.1109/EMR.2023.3285098

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model* (Vol. 18). John Wiley & Sons.

Chien, S.-Y., Sycara, K., Liu, J.-S., & Kumru, A. (2016). Relation between Trust Attitudes Toward Automation, Hofstede's Cultural Dimensions, and Big Five Personality Traits. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 841–845. https://doi.org/10.1177/1541931213601192

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: Concepts, evolving themes, a model. *International Journal of Human-Computer Studies*, *58*(6), 737–758.

Cummings, M. L., & Clare, A. S. (2015). Holistic modelling for human-autonomous system interaction. *Theoretical Issues in Ergonomics Science*, *16*(3), 214–231. https://doi.org/10.1080/1463922X.2014.1003990

de Visser, E. J., Peeters, M. M. M., Jung, M. F., Kohn, S., Shaw, T. H., Pak, R., & Neerincx, M. A. (2020). Towards a Theory of Longitudinal Trust Calibration in Human–Robot Teams. *International Journal of Social Robotics*, *12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

De Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. *International Journal of Human-Computer Studies*, *58*(6), 719–735.

Dekker, S., & Hollnagel, E. (2004). Human factors and folk models. *Cognition, Technology & Work*, *6*(2), 79–86. https://doi.org/10.1007/s10111-003-0136-9

Donmez, B., Boyle, L. N., Lee, J. D., & McGehee, D. V. (2006). Drivers' attitudes toward imperfect distraction mitigation strategies. *Transportation Research Part F: Traffic Psychology and Behaviour*, *9*(6), 387–398.

Emzivat, Y., Ibanez-Guzman, J., Martinet, P., & Roux, O. H. (2017). Dynamic driving task fallback for an automated driving system whose ability to monitor the driving environment has been compromised. *2017 IEEE Intelligent Vehicles Symposium (IV)*, 1841–1847. https://ieeexplore.ieee.org/abstract/document/7995973/

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, *42*(6), 1585–1593.

Forrester, J. W. (1987). Lessons from system dynamics modeling. *System Dynamics Review*, *3*(2), 136–149. https://doi.org/10.1002/sdr.4260030205

Forrester, J. W. (1997). Industrial Dynamics. *Journal of the Operational Research Society*, *48*(10), 1037–1041. https://doi.org/10.1057/palgrave.jors.2600946

Gambardella, P. J., Polk, D. E., Lounsbury, D. W., & Levine, R. L. (2017). A co-flow structure for goal-directed internal change: Co-flow Structure with Goal-Directed, Internal Change. *System Dynamics Review*, *33*(1), 34–58. https://doi.org/10.1002/sdr.1574

Gao, F., Clare, A. S., Macbeth, J. C., & Cummings, M. L. (2013). *Modeling the Impact of Operator Trust on Performance in Multiple Robot Control*. 7.

Gao, J., & Lee, J. D. (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), 943–959.

Ghaffarzadegan, N., Xue, Y., & Larson, R. C. (2017). Work-education mismatch: An endogenous theory of professionalization. *European Journal of Operational Research*, *261*(3), 1085–1097. https://doi.org/10.1016/j.ejor.2017.02.041

Goldberg, L. R., Johnson, J. A., Eber, H. W., Hogan, R., Ashton, M. C., Cloninger, C. R., & Gough, H. G. (2006). The international personality item pool and the future of public-domain personality measures. *Journal of Research in Personality*, *40*(1), 84–96.

Guastello, S. J. (2017). Nonlinear dynamical systems for theory and research in ergonomics. *Ergonomics*, *60*(2), 167–193. https://doi.org/10.1080/00140139.2016.1162851

Guo, Y., & Yang, X. J. (2021). Modeling and Predicting Trust Dynamics in Human-Robot Teaming: A Bayesian Inference Approach. In *INTERNATIONAL JOURNAL OF SOCIAL ROBOTICS* (Vol. 13, Issues 8, SI, pp. 1899–1909). SPRINGER. https://doi.org/10.1007/s12369-020-00703-3

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A Meta-Analysis of Factors Affecting Trust in Human-Robot Interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving Trust in Robots: Specification Through Sequential and Comparative Meta-Analyses. *Human Factors*, *63*(7), 1196–1229. https://doi.org/10.1177/0018720820922080

Hancock, P. A., & Szalma *, J. L. (2004). On the relevance of qualitative methods for ergonomics. *Theoretical Issues in Ergonomics Science*, *5*(6), 499–506. https://doi.org/10.1080/14639220412331303391

Hartwich, F., Witzlack, C., Beggiato, M., & Krems, J. F. (2019). The first impression counts—A combined driving simulator and test track study on the development of trust and acceptance of highly automated driving. In *TRANSPORTATION RESEARCH PART F-TRAFFIC PSYCHOLOGY AND BEHAVIOUR* (Vol. 65, pp. 522–535). ELSEVIER SCI LTD. https://doi.org/10.1016/j.trf.2018.05.012

Hoff, K. A., & Bashir, M. (2015). Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Hollnagel, E. (1993). Requirements for dynamic modelling of man-machine interaction. *Nuclear Engineering and Design*, *144*(2), 375–384. https://doi.org/10.1016/0029-5493(93)90153-Z

Hollnagel, E. (2002). Time and time again. *Theoretical Issues in Ergonomics Science*, *3*(2), 143–158. https://doi.org/10.1080/14639220210124111

Hu, W.-L., Akash, K., Reid, T., & Jain, N. (2018). Computational modeling of the dynamics of human trust during human–machine interactions. *IEEE Transactions on Human-Machine Systems*, *49*(6), 485–497.

Itoh, M., Abe, G., & Tanaka, K. (1999). Trust in and use of automation: Their dependence on occurrence patterns of malfunctions. *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No. 99CH37028)*, *3*, 715–720. https://ieeexplore.ieee.org/abstract/document/823316/

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye* (pp. 573–605). Elsevier. https://www.sciencedirect.com/science/article/pii/B9780444510204500311

Jagacinski, R. J., & Flach, J. M. (2018). *Control theory for humans: Quantitative approaches to modeling performance*. CRC press.

John, O. P., & Srivastava, S. (1999). *The Big-Five trait taxonomy: History, measurement, and theoretical perspectives*. http://www.personality-project.org/revelle/syllabi/classreadings/john.pdf

Jonker, C. M., & Treur, J. (1999). Formal Analysis of Models for the Dynamics of Trust Based on Experiences. In F. J. Garijo & M. Boman (Eds.), *Multi-Agent System Engineering* (Vol. 1647, pp. 221–231). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-48437-X_18

Kim, P. H., Dirks, K. T., & Cooper, C. D. (2009). The Repair of Trust: A Dynamic Bilateral Perspective and Multilevel Conceptualization. *Academy of Management Review*, *34*(3), 401–422. https://doi.org/10.5465/amr.2009.40631887

Kohn, S. C., De Visser, E. J., Wiese, E., Lee, Y.-C., & Shaw, T. H. (2021). Measurement of Trust in Automation: A Narrative Review and Reference Guide. *Frontiers in Psychology*, *12*.

Kongsberg. (2023). *K-Sim—Maritime Simulation—Education, Training and Studies*. https://www.kongsbergdigital.com/resources

Körber, M. (2019). Theoretical Considerations and Development of a Questionnaire to Measure Trust in Automation. In S. Bagnara, R. Tartaglia, S. Albolino, T. Alexander, & Y. Fujita (Eds.), *Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018)* (Vol. 823, pp. 13–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-96074-6_2

Kraus, J., Scholz, D., Messner, E.-M., Messner, M., & Baumann, M. (2020). Scared to Trust? – Predicting Trust in Highly Automated Driving by Depressiveness, Negative Self-Evaluations and State Anxiety. *Frontiers in Psychology*, *10*, 2917. https://doi.org/10.3389/fpsyg.2019.02917

Kugler, P. N., & Turvey, M. T. (2015). *Information, natural law, and the self-assembly of rhythmic movement*. Routledge.

Lane, D. C. (2000). Should system dynamics be described as a 'hard' or 'deterministic' systems approach? *Syst. Res.*, 20.

Lane, D. C., & Schwaninger, M. (2008). Theory building with system dynamics: Topic and research contributions. *Systems Research and Behavioral Science*, *25*(4), 439–445. https://doi.org/10.1002/sres.912

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

Lee, J., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, *40*(1), 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80.

Lewicki, R. J., & Brinsfield, C. (2011). Framing trust: Trust as a heuristic. *Framing Matters: Perspectives on Negotiation Research and Practice in Communication*, 110–135.

Lewis, J. D., & Weigert, A. J. (2012). The social dynamics of trust: Theoretical and empirical research, 1985-2012. *Social Forces*, *91*(1), 25–31.

Lewis, M., Sycara, K., & Walker, P. (2018). The Role of Trust in Human-Robot Interaction. *Studies in Systems, Decision and Control*, *117*, 135–159. https://doi.org/10.1007/978-3-319-64816-3_8

Lu, Y., & Sarter, N. (2019). Eye Tracking: A Process-Oriented Method for Inferring Trust in Automation as a Function of Priming and System Reliability. *IEEE Transactions on Human-Machine Systems*, *49*(6), 560–568. https://doi.org/10.1109/THMS.2019.2930980

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

Manzey, D., Reichenbach, J., & Onnasch, L. (2012). Human performance consequences of automated decision aids: The impact of degree of automation and system experience. *Journal of Cognitive Engineering and Decision Making*, *6*(1), 57–87.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734.

McCarthy, J. T., Hocum, C. L., Albright, R. C., Rogers, J., Gallaher, E. J., Steensma, D. P., Gudgell, S. F., Bergstralh, E. J., Dillon, J. C., & Hickson, L. J. (2014). Biomedical system dynamics to improve anemia control with darbepoetin alfa in long-term hemodialysis patients. *Mayo Clinic Proceedings*, *89*(1), 87–94. https://www.sciencedirect.com/science/article/pii/S0025619613009300

McCrae, R. R., & John, O. P. (1992). An Introduction to the Five-Factor Model and Its Applications. *Journal of Personality*, *60*(2), 175–215. https://doi.org/10.1111/j.1467-6494.1992.tb00970.x

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors*, *50*(2), 194–210.

Moray, N., & Inagaki, T. (1999). Laboratory studies of trust between humans and machines in automated systems. *Transactions of the Institute of Measurement and Control*, *21*(4–5), 203–211.

Moray, N., Inagaki, T., & Itoh, M. (2000). Adaptive automation, trust, and self-confidence in fault management of time-critical tasks. *Journal of Experimental Psychology: Applied*, *6*(1), 44.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. *International Journal of Man-Machine Studies*, *27*(5–6), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

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

Muir, B. M., & Moray, N. (1996). Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, *39*(3), 429–460. https://doi.org/10.1080/00140139608964474

Øvergård, K. I., Nielsen, A. R., Nazir, S., & Sorensen, L. J. (2015). Assessing Navigational Teamwork Through the Situational Correctness and Relevance of Communication. *Procedia Manufacturing*, *3*, 2589–2596. https://doi.org/10.1016/j.promfg.2015.07.579

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410.

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

Poornikoo, M., & Øvergård, K. I. (2023). Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation. *Theoretical Issues in Ergonomics Science*, *0*(0), 1–37. https://doi.org/10.1080/1463922X.2023.2233591

Pop, V. L., Shrewsbury, A., & Durso, F. T. (2015). Individual Differences in the Calibration of Trust in Automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *57*(4), 545–556. https://doi.org/10.1177/0018720814564422

Porter, D., McAnally, M., Bieber, C., Wojton, H., & Medlin, R. (2020). Trustworthy autonomy: A roadmap to assurance-Part 1: System effectiveness. *Inst. Def. Anal., Alexandria, VA, USA, Tech. Rep. AD1131283*, *20*. https://apps.dtic.mil/sti/citations/AD1131283

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95.

Riley, V. A. (1994). *Human use of automation*. University of Minnesota.

Rodriguez Rodriguez, L., Bustamante Orellana, C. E., Chiou, E. K., Huang, L., Cooke, N., & Kang, Y. (2023). A review of mathematical models of human trust in automation. *Frontiers in Neuroergonomics*, *4*. https://www.frontiersin.org/articles/10.3389/fnrgo.2023.1171403

Sanders, T., Oleson, K. E., Billings, D. R., Chen, J. Y. C., & Hancock, P. A. (2011). A Model of Human-Robot Trust: Theoretical Model Development. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *55*(1), 1432–1436. https://doi.org/10.1177/1071181311551298

Sastry, A. (1998). Archetypal self-reinforcing structures in organizations. *Proceedings of the 16th International Conference of the System Dynamics Society, Quebec*. https://proceedings.systemdynamics.org/1998/PROCEED/00003.PDF

Sastry, S. (2013). *Nonlinear systems: Analysis, stability, and control* (Vol. 10). Springer Science & Business Media. https://www.google.com/books?hl=no&lr=&id=j_PiBwAAQBAJ&oi=fnd&pg=PR7&dq=Sastry,+1997+system+dynamics&ots=rCcJrKEF_J&sig=vMrUqJfxh_CaKw7vOTS35ENaOFE

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-Analysis of Factors Influencing the Development of Trust in Automation: Implications for Understanding Autonomy in Future Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 377–400. https://doi.org/10.1177/0018720816634228

Schwaninger, M., & Grösser, S. (2008). System dynamics as model-based theory building. *Systems Research and Behavioral Science*, *25*(4), 447–465. https://doi.org/10.1002/sres.914

Sheridan, T. B. (2019). Extending Three Existing Models to Analysis of Trust in Automation: Signal Detection, Statistical Parameter Estimation, and Model-Based Control. *Human Factors*, *61*(7), 1162–1170. https://doi.org/10.1177/0018720819829951

Sterman, J. D. (2000). Business Dynamics, S. *Massachusetts: Jeffrey J. Shelstad*, *196*, 199–201.

Sterman, J. D., & Wittenberg, J. (1999). Path Dependence, Competition, and Succession in the Dynamics of Scientific Revolution. *Organization Science*, *10*(3), 322–341. https://doi.org/10.1287/orsc.10.3.322

Sweetser, A. (1999). A comparison of system dynamics (SD) and discrete event simulation (DES). *17th International Conference of the System Dynamics Society*, 20–23. https://proceedings.systemdynamics.org/1999/PAPERS/PARA78.PDF

Szalma, J. L., & Taylor, G. S. (2011). Individual differences in response to automation: The five factor model of personality. *Journal of Experimental Psychology: Applied*, *17*(2), 71–96. https://doi.org/10.1037/a0024170

Tenhundfeld, N., Demir, M., & de Visser, E. (2022). Assessment of Trust in Automation in the "Real World": Requirements for New Trust in Automation Measurement Techniques for Use by Practitioners. *Journal of Cognitive Engineering and Decision Making*, 15553434221096261.

Towill, D. R. (1993). System dynamics—background, methodology, and applications. Part 1: Background and methodology. *Computing & Control Engineering Journal*, *4*(5), 201–208.

Van de Ven, A. H. (2007). *Engaged scholarship: A guide for organizational and social research*. Oxford University Press on Demand.

Vehovar, V., Toepoel, V., & Steinmetz, S. (2016). *Non-probability sampling* (Vol. 1). The Sage handbook of survey methods. https://www.google.com/books?hl=no&lr=&id=g8OMDAAAQBAJ&oi=fnd&pg=PA329&dq=non+probability+sampling+methods&ots=DAmGlzY0mP&sig=RZ5EtgaS-XK5aqyUH2fJremTU_8

Walliser, J. C. (2011). *Trust in automated systems the effect of automation level on trust calibration* [PhD Thesis, Monterey, California. Naval Postgraduate School]. https://www.researchgate.net/profile/James-Walliser/publication/235181550_Trust_in_Automated_Systems_The_Effect_of_Automation_Level_on_Trust_Calibration/links/6330564f6063772afd8fe088/Trust-in-Automated-Systems-The-Effect-of-Automation-Level-on-Trust-Calibration.pdf

Williams, T. M. (Ed.). (1997). *Managing and Modelling Complex Projects* (Vol. 17). Springer Netherlands. https://doi.org/10.1007/978-94-009-0061-5

Wittenberg, J. (1992). On the very idea of a system dynamics model of Kuhnian science. *System Dynamics Review*, *8*(1), 21–33. https://doi.org/10.1002/sdr.4260080103

Xu, A., & Dudek, G. (2015). Optimo: Online probabilistic trust inference model for asymmetric human-robot collaborations. *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 221–228.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017). Evaluating Effects of User Experience and System Transparency on Trust in Automation. *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction - HRI '17*, 408–416. https://doi.org/10.1145/2909824.3020230

Yang, X. J., Wickens, C. D., & Hölttä-Otto, K. (2016). How users adjust trust in automation: Contrast effect and hindsight bias. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 196–200. https://doi.org/10.1177/1541931213601044

TABLE 2. Compared means Paired Samples Test, paired differences in time 1 and time 2

	Time 1-Pre-Error (N=28)			Time 2 -Post-Error (N=28)			Paired Differences				
	M	*SD*	*95% CI*	*M*	*SD*	*95% CI*	M_{diff}	*SD*	*95% CI*	*t*	*Cohen's d*
REL	3.46	0.373	[3.32, 3.60]	2.69	0.694	[2.49, 3.01]	0.77	0.616	[0.53, 1.01]	6.651	1.25
TRU	3.68	0.629	[3.40, 3.87]	3.02	1.018	[2.70, 3.46]	0.66	0.82	[0.344, 0.98]	4.279	0.80
Temporal Metrics											
TDF	44963.11	38535.01	[28916, 57216]	101045.39	69226.59	[71718, 122764]	-56082.286	55529.778	[-77614, -34550]	-5.34	1.01
ADF	322.32	69.08	[294.12, 354.42]	390.96	107.84	[346.79, 426.01]	-68.643	108.798	[-110, -26]	-3.33	0.631
Spatial Metrics											
AMS	5.71	0.87	[5.38, 6.03]	5.5	0.90	[5.26, 5.92]	.13964	.59624	[-.09, .37]	1.23	0.234
TAS	4190	1260.82	[3695, 4615]	5162.54	1437.54	[4595, 5639]	-972.53	1267.43	[-1463, -481]	-4.06	0.767
Count Metrics											
NV	56.61	27.87	[45.6, 66.26]	113.04	43.56	[95.19, 127.41]	-56.429	36.839	[-70, -42]	-8.10	1.532
NS	51.54	56.17	[28.10, 69.36]	86.64	56.35	[61.31, 103.62]	-35.107	48.293	[-53, -16]	-3.84	0.727
NF	133.5	102.70	[91.10, 166.5]	253.11	136.54	[195.14, 295.99]	-119.607	111.021	[-162, -76]	-5.70	1.07

Notes. REL = perceived reliability; TRU = trust in automation; TDF = total duration of fixation; ADF = average duration of fixation; AMS = average amplitude of saccades; TAS = total amplitude of saccades; NV = number of visits; NS = number of saccades; NF = number of fixations; M = mean; SD = standard deviation; 95% CI = 95% confidence interval, M_{diff} = mean difference; t = t-statistics

TABLE 3. Means, Standard Deviation, Internal Consistency (Cronbach's α), and Correlations of all traits and Trust in Automation in time 1 and time 2.

Variable	α	*M*	*SD*	1	2	3	4	5	6	7	8	9	10
1. REL t_1	.40	3.46	0.39										
2. TRU t_1	.77	3.68	0.63	.65** [.37, .82]									
3. REL t_2	.78	2.69	0.68	.44* [.08, .70]	.55** [.22, .77]								
4. TRU t_2	.93	3.02	1.02	.32 [-.06, .62]	.60** [.29, .80]	.79** [.59, .90]							
5. dTRU	_	-0.62	0.82	-.09 [-.45, .29]	-.06 [-.43, .32]	.51** [.18, .74]	.74** [.51, .87]						
6. dREL	_	-0.77	0.62	-.15 [-.49, .24]	.20 [-.19, .53]	.83** [.65, .92]	.66** [.39, .83]	.62** [.33, .81]					
7. E	.82	24.79	6.59	.12 [-.27, .47]	.02 [-.35, .39]	.10 [-.29, .45]	.26 [-.12, .58]	.29 [-.09, .60]	.03 [-.34, .40]				
8. A	.75	28.32	5.89	.12 [-.27, .47]	-.02 [-.39, .36]	-.14 [-.49, .24]	-.11 [-.46, .28]	-.07 [-.43, .31]	-.23 [-.56, .15]	.35 [-.03, .64]			
9. C	.69	25.11	5.82	.08 [-.30, .44]	.35 [-.02, .64]	.13 [-.25, .48]	.14 [-.25, .49]	-.04 [-.41, .34]	.09 [-.29, .45]	-.13 [-.48, .25]	-.02 [-.39, .36]		
10.N	.78	26.86	5.87	.03 [-.34, .40]	.01 [-.36, .38]	.14 [-.25, .49]	.06 [-.32, .43]	.04 [-.34, .41]	.13 [-.25, .48]	-.05 [-.41, .33]	.13 [-.26, .48]	.42* [.05, .69]	
11.O	.73	27.18	5.12	.15 [-.24, .49]	.34 [-.04, .63]	.06 [-.32, .42]	.24 [-.15, .56]	.09 [-.29, .45]	-.03 [-.40, .35]	.01 [-.37, .38]	.24 [-.14, .56]	.37 [-.01, .65]	.04 [-.34, .41]

Note. REL = perceived reliability; TRU = trust in automation; dTRU = change in trust in automation, dREL = change in perceived reliability E = extraversion; A = agreeableness; C = conscientiousness; N= neuroticism; O = openness; t_1 = measurement at time 1, t_2 = measurement at time 2, α = Cronbach's alpha; M = mean; SD = Standard deviation, Values in square brackets indicate the 95% confidence interval for each correlation, * indicates $p < .05$. ** indicates $p < .01$.