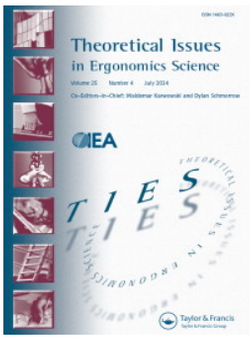

Theoretical Issues in Ergonomics Science

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/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 (2024) Model evaluation in human factors and ergonomics (HFE) sciences; case of trust in automation, Theoretical Issues in Ergonomics Science, 25:4, 416-452, 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
2024, VOL. 25, NO. 4, 416–452
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

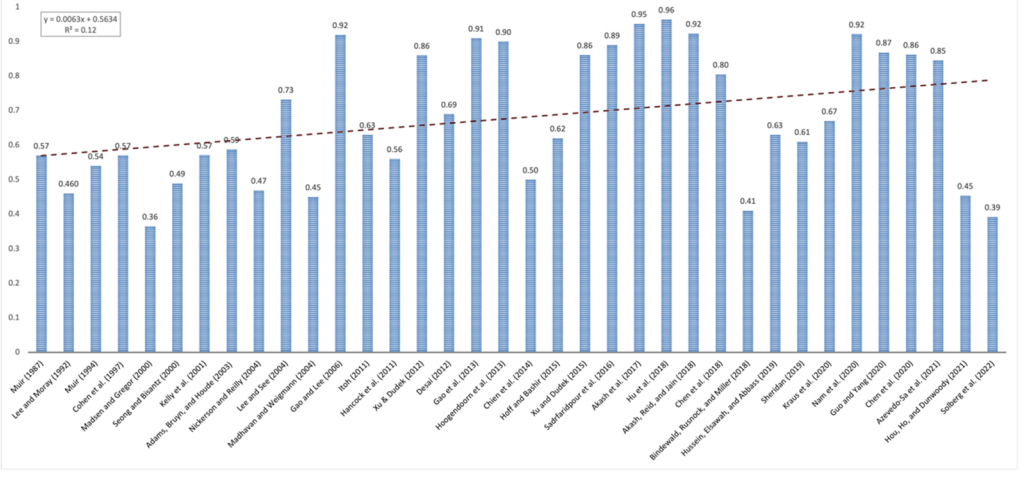

Figure 1. TiA model scores over time.

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.