Taylor & Francis
Taylor & Francis Group

Journal of Simulation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/tjsm20

Model credibility revisited: Concepts and considerations for appropriate trust

Levent Yilmaz & Bo Liu

To cite this article: Levent Yilmaz & Bo Liu (2020): Model credibility revisited: Concepts and considerations for appropriate trust, Journal of Simulation, DOI: 10.1080/17477778.2020.1821587

To link to this article: https://doi.org/10.1080/17477778.2020.1821587

JOURNAL OF SIMULATION
https://doi.org/10.1080/17477778.2020.1821587

Model credibility revisited: Concepts and considerations for appropriate trust

Levent Yilmaz and Bo Liu

Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, USA

ABSTRACT

The increasing reliance of modern science in computer simulation demands appropriate trust in simulation models for credible results. Because of its foundations in operations research, model credibility is conventionally viewed from the lens of numerical and transformational accuracy. However, the exploratory use of models in scientific discovery, causal explanation, and strategic decision-making render this view incomplete. Recognising the significance of the cognitive interests of model users and the context-sensitive, adaptive nature of building confidence in scientific models, we characterise credibility as trust. Appropriate and justifiable trust is conceptualised as a dynamic, cognitive construct that evolves through interactive, experiential learning. Following the delineation of the dimensions and attributes of trust, conceptual foundations of a dynamic trust model, including alternative measurement strategies, are proposed. Guidelines for trustable models are elaborated to provide a basis for exploiting synergies between the cognitive models of trust and model evaluation.

ARTICLE HISTORY
Received 25 November 2019
Accepted 7 September 2020

KEYWORDS
Model credibility; trust; trustable model; validation; cognitive model; credibility assessment

1. Introduction

Contemporary science increasingly relies on computational models for advancing knowledge. As simulation models facilitate exploration of explanations for scientific phenomena, and the discovery of robust strategies under uncertainty, instilling confidence in their behaviour is becoming paramount across a broad range of application categories (Onggo et al., 2019). Among these categories include the prediction of system behaviour (Rossiter, 2017), inferring the causal mechanisms of complex systems by exploring possibilities (Larsen et al., 2014), supporting robust decision-making under uncertainty (Davis et al., 2018), training in virtual environments, intelligent tutoring in education, controlling cyber-physical systems (E. A. Lee, 2008; Tolk et al., 2018), and diagnosis of behaviour when a model reports its state (Yilmaz, 2004). Although these categories are not exhaustive, they identify areas that require research for a more in-depth understanding of model credibility while taking into consideration the characteristics of trustworthiness and competence in each domain.

Model developers view credibility traditionally from the perspective of verification and validation that evaluate a model throughout its development lifecycle (Balci, 1986; Sargent, 1983). Validation aims to substantiate the similarity of a model's behaviour to the behaviour of the referent for the intended purposes within its domain of application. As the objectives of models, their intended applications, and the nature of problems evolve, credibility assessment strategies need to adapt as well. According to a recent study by Davis et al. (2018), scientific problems, especially in the context of natural, social, and behavioural sciences, exhibit the inherent characteristics of Complex Adaptive Systems (CAS). Due to the difficulty of predicting CAS behaviour, the study (Davis et al., 2018) posits that it is necessary to assess credibility by leveraging criteria across multiple dimensions, including causal explanation, prediction, exploratory analysis, and description.

Similarly, Gelfert (2019) highlights the use of models as exploration tools as well as instruments in the development of *credible worlds* to gain intuition about scientific phenomena in the absence of established theory and data. Such uses of models are offered as examples to illustrate the limitations of conventional empirical similarity measures. Instead, a model's credibility is viewed not merely as a function of its features but also from the perspective and cognitive interests of a scientist (Knuuttila, 2005). These recent developments suggest the need for a better understanding of the processes, principles, and methods for instilling confidence in models continuously and dynamically as scientists further their inquiry. Such understanding facilitates the development of computational strategies and tools for improving transparency and enabling justification of credibility through experiential learning over a broad range of simulation experiments.

According to the Merriam Webster dictionary (Merriam-Webster, 2019), *credibility* is the quality of inspiring belief or the quality of being convincing or believable. In this paper, credibility is construed as a perceived measure of *believability*. As a perceived quality, it has multiple dimensions that are

CONTACT Levent Yilmaz yilmaz@auburn.edu Department of Computer Science and Software Engineering, Auburn University, Auburn, AL, USA

concurrently monitored to generate an evolving judgement. To this end, we examine the key concepts and terms that relate to credibility and then analyse the extant literature to discern emergent conceptual patterns. Based on this analysis and the recognition of the inherent challenges associated with the complexity and diversity of scientific problems, we propose a generic framework for the assessment of model credibility from the cognitive perspective of trust.

2. Background

As the function and purpose of simulation models continue to expand, their use in computational science reveals the limitations of the traditional systems engineering view of credibility assessment (Gelfert, 2019; Onggo et al., 2019). The conventional accuracy-driven perspective of credibility stems from the need for assessing whether models can serve as surrogates for well-defined systems. However, when models are used for exploring plausible explanations, as proofs-of-concepts, or as components in support systems for strategic decision-making under uncertainty, empirical accuracy is only part of the broader assessment scheme. It may not even be applicable in the absence of data and for systems with little prior (Davis et al., 2018; Gelfert, 2016; Onggo et al., 2019; Young, 1983). Recognising the limitations of traditional credibility assessment methods for such systems, Onggo et al. (2019) highlight the role that trust can play in the acceptability of simulation models.

A recent study (Harper et al., 2020) provides a synthesis of literature and outlines the facets of trust in relation to aspects pertaining to model developers, stakeholders, and model representation. The authors note the scarcity of research on trustworthiness with respect to simulation modelling while acknowledging the considerable amount of existing work in the information systems and management fields. The primary focus in the management domain is on interpersonal,team and organisational levels of trust (Ebert, 2009; Fulmer & Gelfand, 2012) while the information systems research concentration is on technology acceptance and adoption (Wang et al., 2015).

To map the state of research and discern further growth channels in model credibility assessment, we use the keyword co-occurrence network-based systematic review method (Radhakrishnan et al., 2017). Keyword co-occurrence networks (Chen, 2006; Peters & van Raan, 1993) facilitate knowledge mapping and to conduct systematic reviews of scientific literature. In such networks, nodes represent keywords, and a link that connects a pair of keywords depicts their co-occurrence in one or more articles. The patterns and strengths of links represent cumulative knowledge and help detect clusters of and associations between fields. For our analysis, we used the following keywords: *model, simulation, credibility*, and *computational model*. The *VosViewer* software (Van Eck & Waltman, 2010) is used over the Web of Science database, and the keywords are selected to narrow our analysis over the papers that focus on the credibility assessment of dynamic computational models in simulation studies across a broad range of disciplines.

The keyword co-occurrence network analysis of over 450 papers in the extant literature indicates distinct focus areas in the assessment and evaluation of credibility. Figure 1 depicts the relational co-occurrence strength of keywords in published articles related to simulation model credibility across multiple disciplines. According to the emergent network structure, the term *credibility* associates with articles across multiple clusters, which reveal groupings of articles within specific disciplinary areas. The articles that reference the term *credibility* in the context of the M&S cluster often use it in conjunction with the terms *validation, simulation, verification*, and *uncertainty*. The computing/automation domain uses the term credibility along with the terms *algorithm, design, network*, and *risk*. Common terms that frequently co-occur with credibility in the domain of social and behavioural sciences are *trust, behaviour*, and *management*. Other clusters based on keyword association include decision-making, communication-media studies, credibility theory, and marketing. Each one of these domains include references to *risk* or *trust* in the context of *credibility*. Yet, within the traditional M&S domain, *risk* and *trust* are rarely used in relation to credibility.

2.1. Verification and validation of predictive models

In Modelling & Simulation, the conventional interpretation of credibility relies on the concepts of model verification and validation (Balci, 1986), and its assessment is viewed as a continuous activity throughout the lifecycle of a simulation study. Balci (2015) suggests various guidelines and indicators to assess the credibility of a simulation model across multiple phases, including problem formulation, model design, implementation, and experimentation. Verification involves the transformational accuracy of models as they are refined and translated into increasingly concrete representations, resulting in the implementation of the conceptual model in a simulation program. Although verification is widely considered as the process of assuring correct implementation of a conceptual model, in practice, model verification is a continuous activity. The correctness, consistency, and completeness of a model design against a conceptual model is as much part of this multi-step

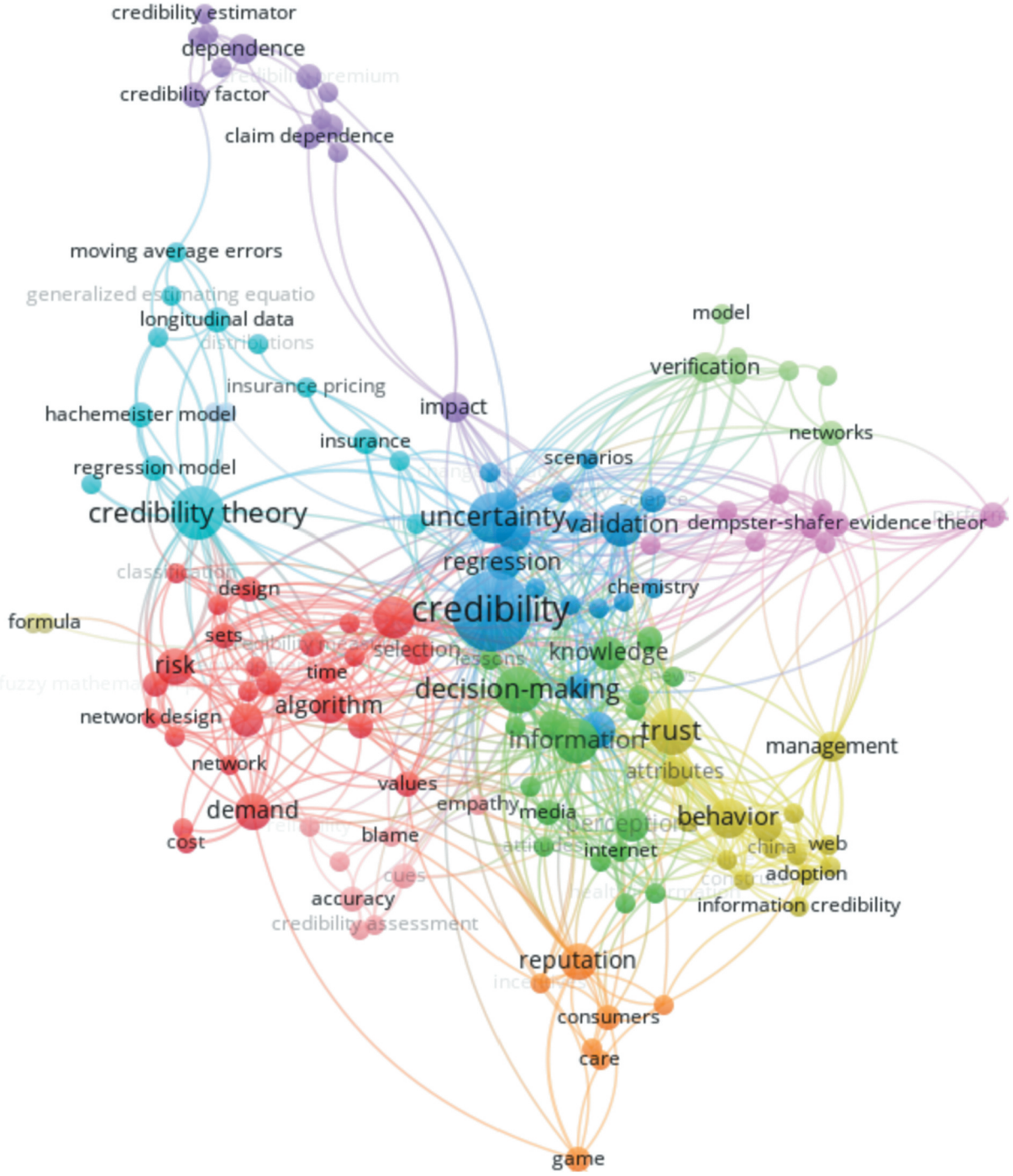


Figure 1. Credibility as a transdisciplinary concept.

verification process as the evaluation of a design model against the implementation.

The validation aspect of evaluation aims to substantiate that the model has a satisfactory range of accuracy within its domain of applicability from the perspective of its intended purpose and uses (Sargent, 2013; Sargent & Balci, 2017). The ability to predict and reproduce a system's behaviour under well-defined conditions is a critical aspect of model validation. Therefore, validation is viewed from the lens of the context of inquiry and the objectives of the model user. The reliability of simulations and the validity of models are also viewed as a matter of trust (Casti, 1997), albeit still from the perspective of accuracy. However, accuracy-driven validation requires the availability of empirical data, as well as an explicit and precise description of the expected system behaviour. Such specifications help determine whether the observed similarities between the system and the simulated model are for the right reasons; that is, whether the underlying causal processes are also sufficiently similar (Yilmaz, 2006). Acknowledging the limitations of using accuracy as an exclusive metric for credibility, Johnson (2000) also suggests the

participation of stakeholders in the context of socio-technical systems.

2.2. Exploratory use of models in science

The functions and purposes of models extend beyond their ability to reproduce the behaviour of a target system measured by metrics such as similarity, empirical fitness, distance, and statistical accuracy. Models rely on simplifying metaphorical abstractions and representational tools that are pragmatically adequate to study the behavioural regularities of a referent. In constructing such credible worlds (Gelfert, 2019), a model developer may not be able to start from the actual real-world system through isolating observable, causally realistic factors. Instead, the linkage between a model and its referent needs to be established subsequently via controlled simulation experiments and experiential learning. Even if a model does not depend on a well-developed theoretical account or formal specification of the target system, it can be inductively concluded that the metaphorical model is credible for the intended purposes.

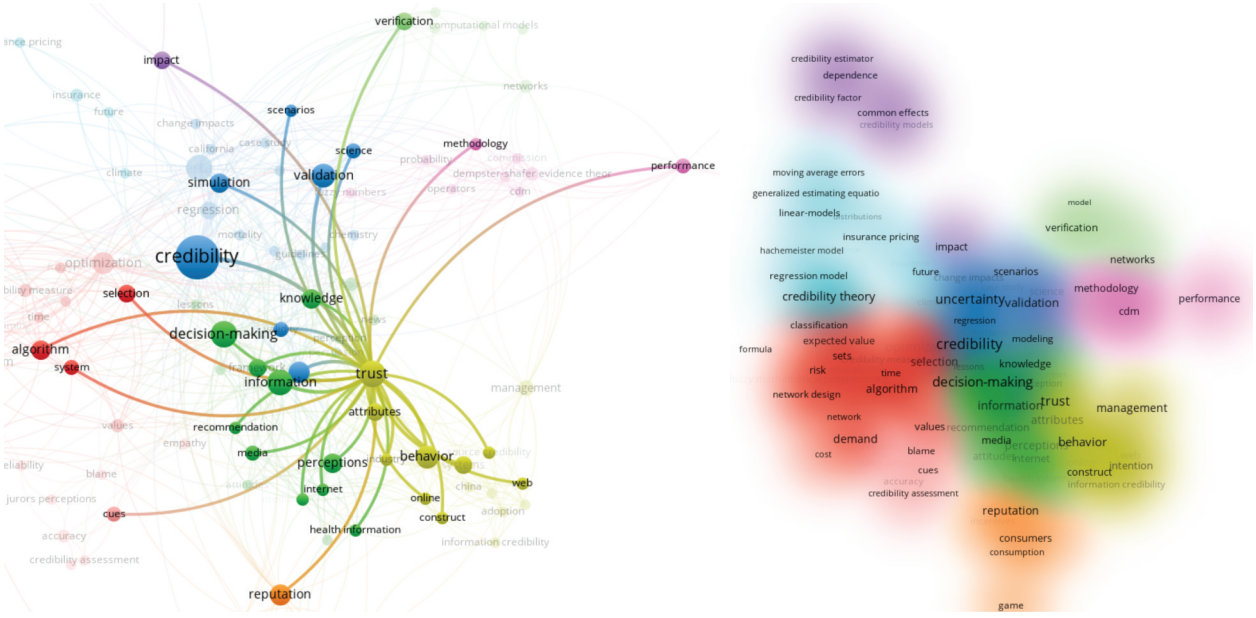
Simulation models are often used to discern plausible explanations for scientific phenomena via exploratory analysis across a range of structural assumptions as well as contextual conditions (Davis et al., 2018; Larsen et al., 2014; Yilmaz, 2004). In this case, whether a model is credible or not is not solely determined based on how faithful its representation to the target system. Instead, its credibility also depends on how promising and fruitful the model is in broadening inquiry and future performance. Hence, in exploratory studies aimed to explain a wide range of phenomena in a specific domain, credibility is often forward-looking. Although such a model may not necessarily serve as a *surrogate* for the referent system, it can *substitute* as a tool for prospective and retrospective (e,g, counterfactual) inquiry about the referent's behaviour (Gelfert, 2016).

3. Characterisation of model credibility as trust

In communication and media studies, the credibility of a source or an interlocutor relies on their *competence* and *trustworthiness*. Human interlocutors are expected to continuously provide reliable information across a wide range of inquiries without deviation from expectations (Self, 2014). Similarly, in Human-Computer Interaction research, (Fogg & Tseng, 1999) highlights both *trust* and *expertise* as the key components of credibility. They argue that credibility is a perceived quality that results from the evaluation of multiple dimensions.

The association between credibility and trust is also evident in keyword co-occurrence analysis of the papers identified in the Web of Science database. Within those papers published in the topical areas of simulation, modelling, and credibility assessment. as shown in Figure 2a, the term *credibility* is used across multiple scientific domains and is associated with a broad range of keywords. In the context of M&S and Computer Engineering, *credibility* is used in conjunction with keywords such as *validation, uncertainty, verification, simulation, algorithm, quality*, *optimisation*, *system*, and *accuracy*.

The papers that are clustered under the social, behavioural, information, and management sciences



(a) Trust

(b) Clustering of Disciplinary Areas

Figure 2. Association of credibility and trust.

use the keywords *trust* and *reputation* frequently along with the term *credibility*. In these publications *trust* co-occurs with *credibility, validation, verification, information, knowledge, decision-making, science, algorithm, media, recommendation, system, simulation, internet, web*, and *behaviour*. Figure 2b illustrates these associations and reveal that the use of *trust* in relation to *credibility* is more common in the context of science than in the context of M&S and engineering. This observation suggests the need for advancing the theory and methodology of M&S to include *trust* as a critical factor in measuring credibility, especially for applications in a broad range of scientific domains, including but not limited to behavioural sciences.

3.1. Credibility as trust

Despite distinct characteristics of the application domains, there are fundamental, overarching principles of credibility as a function of trust and expertise. Credibility, when viewed from the perspective of trust, is a dynamic construct. Models gain trust when they provide results that users find reliable or accurate. However, two models with the same level of empirical accuracy can be perceived to have different levels of trust due to the cognitive interests and goals of the model users. Such interests can evolve and even become emergent based on the information gleaned during the use of the model.

Moreover, if models have multiple stakeholders with different perspectives, as is common in model-driven engineering and science projects, collaborative judgements of credibility and group decision making become integral in credibility assessment. Models lose credibility when they provide incorrect or inaccurate results. Therefore, the provision of correct results should be continuous without exceptions. On the other hand, if model errors can be anticipated and are consistent, and if users develop an understanding of the contextual situations a model is erroneous, the users may be willing to take a controlled risk in using the model. As a result, credibility assessment is a continuous belief formation and revision process. The process entails belief evaluation strategies, which examine both the consequences of and explanations for beliefs. That is, the credibility of the individual features of a model influences the credibility of other features because credible consequences of and explanations for the elements of a model increases the overall credibility of a model.

3.2. Contextual factors

The perceived credibility of a model depends not only on its features and results but also on the context of its use. In unfamiliar scenarios, model users can give higher credence to a model than is warranted. Additional factors can include situations with varying levels of risk and cognitive load. The familiarity of model users with the domain of application can also influence perceived credibility. Users that are familiar with the domain evaluate a model more rigorously, whereas those who are unfamiliar with the underlying background knowledge are likely to view a model credible. Although empirical fitness is an important criterion in credibility assessment, there exist arguments (Parker, 2009) that give adequacy for purpose higher preference as a desirable criterion for many modelling contexts. That is, credibility assessment is tentative and context-dependent because the domain knowledge is often incomplete, or the theory may not be directly applicable in a given situation. Furthermore, because there are trade-offs among generality, precision, and accuracy (Levins, 1966), it is not reasonable to expect a model to be appropriate for all contexts.

3.3. Coherence and credibility

Understanding the causal structure of the background domain knowledge that underlies expected behaviour can increase confidence in simulation outcomes. Therefore, explainability becomes a critical aspect of the credibility assessment of models. According to (Knutti, 2008), trusted model results are those that we can understand the best, and relate to normative conceptual or theoretical frameworks. Therefore, scientists do not find models credible merely because they have empirical adequacy. Credibility attribution depends on whether the model is systematically reliable across a broad range of relevant context and research questions while cohering with the background domain knowledge.

4. Concepts and properties of trust

Trust is a multidisciplinary concept with a broad range of applications. The dictionary definition of trust is "assured reliance on the character, ability, strength, or truth of someone or something" (Merriam-Webster, 2019). In the context of computational agency, Falcone and Castelfranchi (2001) formalise trust as a mental state and present its constructs in terms of competence and predictability. In the context of cooperative relations, trust is viewed as subjective probability that a *trustee* will behave in a way that avoids violation of expectations of a *trustor* under ambiguity (Gambetta, 2000). The role of trust in the production of scientific knowledge and the functioning of the research enterprise is well acknowledged in the epistemology and philosophy of science (Hardwig, 1991). Psychologists define trust as a dynamic concept acquired by learning through interaction based on positive and negative experiences (Rotter, 1980).

In automation, trust is viewed as a mental state that a system facilitates the achievement of expectations in situations characterised by ambiguity and vulnerability (J. D. Lee & See, 2004). As shown in Table 1, trust can be analysed across three primary dimensions. Analytic, affective, and relational constructs play critical roles in the formation and development of trust. Analytic constructs support rational decision-making based on repeated interactions with the trustee (e.g., model) and the cumulative experience over the outcomes (Janani & Manikandan, 2018; S. Marsh & Briggs, 2009). There are also affective mental states such as *expectation, frustration, disappointment, disposition*, and *regret* that influence the evaluation of trust (Falcone & Castelfranchi, 2001). Computational models of such states can improve the effectiveness of logical and analytical strategies by signalling deviations from expectations. Such signals can steer analytical evaluation strategies with limited resources and capacities. Just as emotions serve to redistribute cognitive resources and manage priorities, analogical and relational factors (Self, 2014) can expedite evaluation by relating to similar credible models or contextual factors, especially in the collaborative assessment of credibility.

Table 1. Attributes of trust.

Dimensions	Attributes		Definition
Analytic	belief		mental state, attitude towards model based on capability
	competence	presumed	positive evaluation of quality based on general assumptions
		reputed	based on model use by others
		experienced	based on use by the evaluator
	certainty		degree of confidence
	dependability	reliability	consistency, stability
		availability	accessibility
	utility (risk)		consequence if trust succeeds (fails)
	importance		significance of the problem
	explainability		transparent cause-effect relations
Affective	expectation		desire (goal) for competence
	regret		difference between expected and observed utility/gain
	frustration		invalidation of positive expectation
	forgiveness		tolerance for trust restoration
	disposition	pessimistic	general attitude – worst case view
		optimistic	best case view
		realistic	average case view
Relational	analogy		similarity to successful models
	centrality		degree of influence of the model
	coherence		consistency among model elements coherence with domain knowledge
	reputation		belief and opinions held by others

4.1. Analytic attributes of trust

Analytical components of trust evaluation involve those components that can be logically derived based on observations or empirical evidence.

Belief formation and revision provide a sound framework to evaluate trust from the perspective of decision-making. For instance, Bayesian inference is a popular technique for estimating trust (Janani & Manikandan, 2018). Similarly, fuzzy logic characterises uncertainty and model trust in the context of semantic web and network design. Nilsson (2014) examines what beliefs are and how we acquire and evaluate to develop a coherent knowledge and interpretation of our world. The dynamics of belief formation and update in terms of belief networks based on evidential reasoning can serve as a conceptual framework to develop trust models.

Experience can be used as a proxy measure to determine the expertise level of a model as a function of positive and negative outcomes observed through experimentation. If E^+ is the number of experiments with expected outcomes and E^- denotes the outcomes that deviate from expectations, then

$$\frac{E^+ - E^-}{E^+ + E^-} \tag{1}$$

can be viewed as the degree of experience. Cumulatively, positive and negative experiences, $E^+ + E^-$, indicate the extent to which observations are accumulated to assess *confidence* in the experience measurement for an artefact.

Certainty is a measure that determines the degree of confidence in a specific context. Uncertainty is often related to trust. As a result, general frameworks for reasoning about uncertainty are used to measure trust. For instance, the Dempster-Shafer theory (DST) (Dempster, 2008) provides a basis for defining confidence as a function of belief and plausibility, which are the key constructs of the DST. As uncertainty (U) decreases, the confidence (C) about model behaviour is expected to increase; therefore, $C = 1 - U$ and $U = PL - B$ where plausibility (PL) and belief (B) determine the degree of uncertainty. The consistency of behaviour is also a critical attribute that contributes to trust. In systems engineering, *reliability* and *availability* are often used to assess dependability of a system. By analogy, reliability can be interpreted as consistency (e.g., avoidance of model failures), whereas *availability* indicates the accessibility of models as services over a distributed infrastructure. *Dependability* can be defined by combining these two metrics, suggesting that the model meets the conditions of both metrics.

Expertise or *competence* is another important construct that contributes to trust (Lerch & Prictula, 1989). The measure of competence depends on whether the

model (1) is not previously used to address a specific problem, (2) is used in a broad range of problems, except the current problem, and (3) is known and trusted. When no prior information exists about the model, the only measure is the disposition or perception of the user and the perceived importance of the task that the model aims to address. Such *presumed competence* relies on the general assumptions and prior beliefs of the user towards the model. *Reputed competence* depends on prior use of the model by others and the influence of the model's reputation within a community of users. *Experienced competence* is similar to reputation-based evaluation but relies on the first-hand experience by the model user.

Explainability of a model improves transparency by providing cause-effect relations that contribute to the comprehension of the model (Onggo et al., 2019). When model users gain insight about a model's behaviour and determine that the results are not coincidental but instead as intended and due to expected causal mechanisms, the perceived integrity of the model increases.

4.2. Affective attributes of trust

Human decision-making for trust involves not only analytic but also affective, cognitive processes. The role and significance of affective states in decision-making are well-acknowledged (J. D. Lee & See, 2004). Facilitating appropriate trust in simulation models depends on presenting information about the structure and behaviour of a model in a way compatible with not only analytical processes but also affective states that influence trust. For instance, a trust model that evolves through repeated interactions with the model can be used to measure affective modes such as surprise, frustration, expectation, and disappointment. Formal models that measure states can highlight the root cause of the simulated and emergent emotion to narrow the focus of inquiry and further analysis.

Among the affective states that influence trust, *expectation* is a primary factor, which is critical in providing a basis for prediction. From a cognitive point of view, when the goal of the study has a high degree of subjective significance and importance, anxiety tends to be higher. Expectations can have positive or negative valence. Positive expectations are represented by *hope* whereas *fear* measures the feeling that expects an unfavourable outcome. *Frustration* or disappointment occurs when the positive expectation cannot be validated – the invalidation of a negative expectation results in *relief*.

The degree of trust towards a trustee also depends on the trustor's *disposition*. That is, the evaluation of trust and its estimation relies on the disposition state, which can be *optimistic, pessimistic*, or *realistic*. In the context of distributed AI, trust is described within a spectrum of realism, where pessimists are less likely to be forgiving. S. P. Marsh (1994) describes the trust of an agent i in j (trustee) in a specific context α as

$$T(i,j,\alpha) = T(i,j) \times U(i,j,\alpha) \times IM(i,j)$$

where $T(i,j)$ determines the level of trust in j based on i's disposition. $U(i,j,\alpha)$ characterises the utility gained by i by trusting j, and $IM(i,\alpha)$ estimates the importance of the context or situation α to i. Depending on the disposition state of i, $T(i,j)$ can be assessed differently. Under the *realism* perspective, trust can be considered as the average degree of trust across the foreseeable situations. That is,

$$T(i,j)^R = \frac{1}{|A|}\sum_{\alpha \in A} T(i,j,\alpha)$$

where A denotes the set of all contextual situations in which i interacts with j. An *optimistic* agent takes into account the situation that gives the highest degree of trust and considers it as a typical situation across all scenarios:

$$T(i,j)^O = \max_{\alpha \in A} T(i,j,\alpha)$$

A pessimistic agent takes the opposite view and considers the worst-case scenario as the trust indicator across all situations.

$$T(i,j)^P = \min_{\alpha \in A} T(i,j,\alpha)$$

Considering the relation between trust and willingness to take the risk, an affective state that facilitates making the connection between these two constructs is the concept of *regret*. According to (Luhmann, 1979), "trust is required only if a bad outcome would make you regret your decision". Therefore, regret can serve as a measure for reducing an undesirable experience in the future. Regret is formalised in (S. Marsh & Briggs, 2009) in terms of the difference between the expected and observed utility gain, which is moderated by three factors: the loss, the significance, and affective perception (per) about the utility difference. While loss (lo) and significance (imp) can be determined by functions that measure the utility and importance to the trustor, the affective state can be a challenge to quantify. The resultant expression is

$$Regret(i,j,\alpha) = (U(i,j,\alpha) - U(i,j,\overline{\alpha})) \times F(loimp, per)$$

.

4.3. Relational attributes of trust

In situations that involve a group of stakeholders, collective evaluation of trust requires considering relational attributes of trust. The social interpretation of trust arises in multidisciplinary research projects

where a group of scientists and engineers develops models based on multiple disciplines with different perspectives and evaluation criteria. In such cases, reputation and group decision-making are factored into the overall credibility assessment process. Group decision-making strategies facilitate *collective trust evaluation* so that the views of alternative perspectives and priorities are taken into consideration.

Analogical reasoning can also be useful in the evaluation of model artefacts by demonstrating their similarity to reference models, standards, or observations of phenomena. Similarity can be formalised in terms of weighted feature matching (Weisberg, 2012) between a target and the model under evaluation. Besides, the degree coherence of the assumptions underlying the model with the background theory can provide increased confidence in a model. A model's ability to explain existing evidence while being explained by hypothesised assumptions that also align with evidence facilitates broadening and deepening the support for a model. To this end, theories of cognitive coherence, including explanatory and analogical coherence (Thagard, 2002), can be used to view model constructs embedded within a knowledge network of the discipline. Relational and coherence metrics can assess the degree of relevance and acceptability of a model and its elements within a broader context.

If the components of a model are part of a knowledge network shared by members of a scientific community, standard metrics can be used to assess the importance, relevance, and degree of acceptability. The credibility of such components, in turn, contributes to the credibility of the overall model. For instance, *centrality* or *betweenness centrality* of an element in a network indicates its significance and influence in knowledge representation in a discipline. The number of direct ties associated with a given element facilitates measuring its centrality. On the other hand, betweenness centrality measures how much an element enables directly or indirectly interactions between the nodes that it is connected. The larger the number of nodes that can reach each other through an element in the network, the higher the betweenness of that element. Such elements tend to be reused across broad a range of inquiries that requires the involvement of multiple perspectives, aspects, and resolutions of a system's representation.

5. Dynamics of trust

Trust is a dynamic construct that requires continuous monitoring and evaluation during the incremental and iterative development of models. Such monitoring is critical as the design, and the use of a model can evolve to address a broader range of inquiries while being refined to explore the system of interest at different levels of abstraction. By observing a model's behaviour under controlled settings via simulated experiments, model users can evaluate the degree and appropriateness of trust. Figure 3 illustrates the critical components necessary for context-sensitive trust evaluation. The context includes not only individual belief dispositions towards the model but also social and technical factors associated with the field and the discipline. Each discipline has accepted practices, methods, and domain knowledge that govern the production of new knowledge. The integrity of a model relies on its consistency and coherence with such domain knowledge.

Moreover, the members of a discipline that constitute the field of a domain follow specific norms and criteria for the evaluation of modelling artefacts. The reputation of and trust in individuals and organisational units influence the initial disposition towards a model. Along with the contextual, prior knowledge that is factored into the belief formation process, trust assessment follows a dynamic process governed by objective metrics predicated on empirical and observed behaviour to adjust or reinforce beliefs.

The analytical, affective, and relational attributes highlighted in the previous section provide a basis for trust evolution. An explicit trust model can be used to formalise and explain the data-driven evolution of trust while also providing a basis for updating a model. Model updating is necessary to select those features that prove to be effective in supporting the objectives of the simulation study. Because multiple features compete or that different research questions need alternative model configurations, a multimodel can be used in conjunction with a trust model. When there is uncertainty, the trust model can facilitate proactive experiment management to discern which model features are more effective in a specific situation. The results of experiments across a broad range of situations support an aggregate trust measure for the model.

The mismatch between trust and the capabilities of a model is a significant concern in model use. Supporting appropriate trust to avoid both over trust and unwarranted mistrust is essential for reliable decision-making. If trust exceeds the capabilities of a model, the misuse of a model in irrelevant situations results in flawed conclusions. Model calibration requires a careful balance between capabilities and the extent of trust. Specifically, if there are multiple behavioural regularities and the model successfully generates only a few of such expected behaviour with high accuracy and precision, the model cannot be appropriately trusted across the broader domain of applicability and intended purposes. Such trust is an example of overtrust. Alternatively, mistrust happens when minor discrepancy results in discarding a model

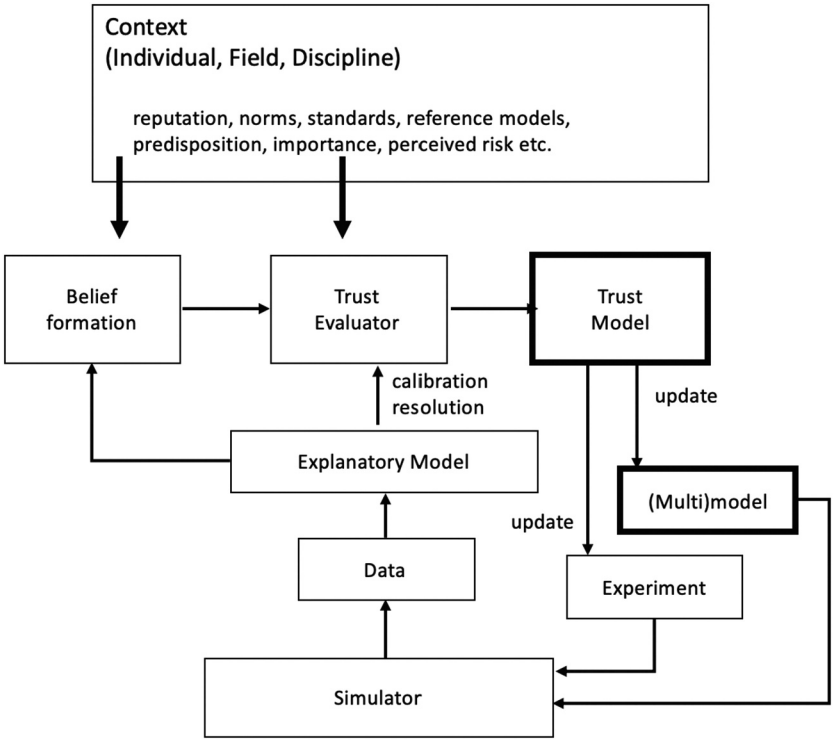


Figure 3. The sketch of a trust assessment process.

despite its demonstrable benefits and capabilities in a broad range of scenarios.

In addition to calibration, model resolution plays a role in assessing the appropriateness of trust by

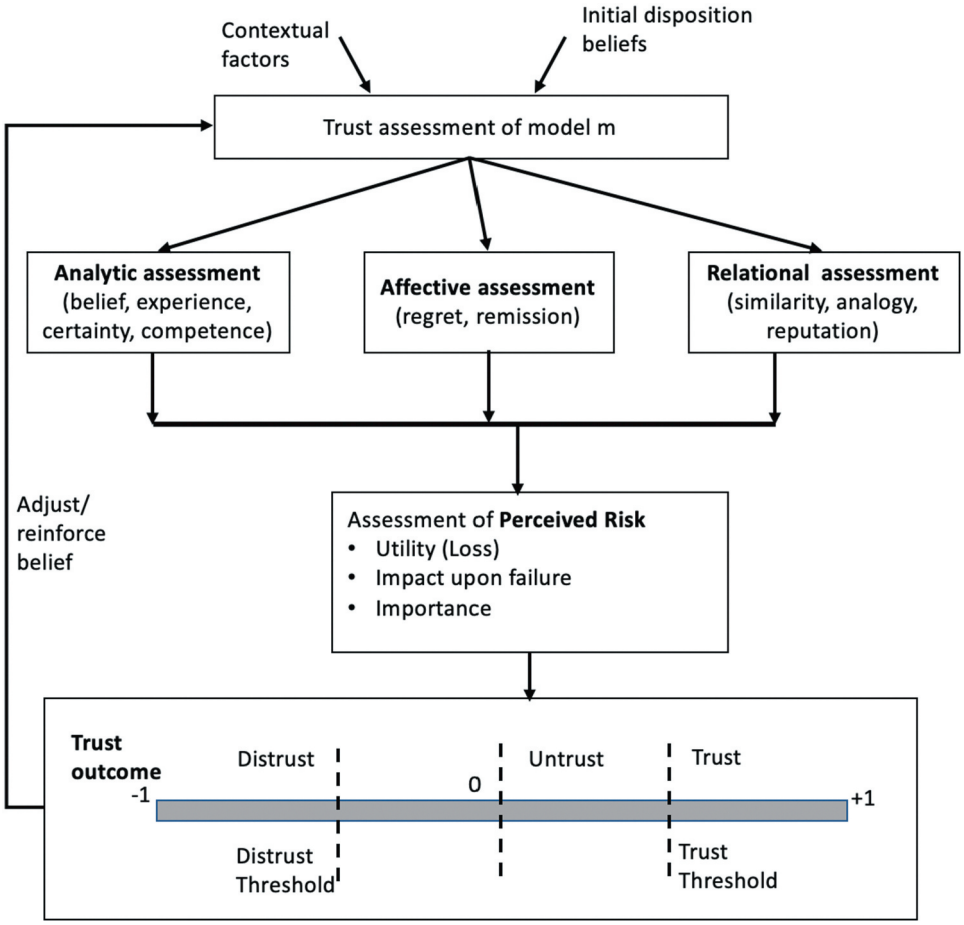


Figure 4. Factors in trust evaluation.

deepening beliefs about a model and its underlying premises. A model gains high credibility if it not only explains observations but is explained or supported by features having high credibility on their own. Therefore, appropriate trust requires broadening and deepening of causal mechanisms that underlie a model, resulting in an incremental and iterative process for building trust. At each step in the process, analytical, affective, and relational elements are considered to update the trust measurement. As shown in Figure 4, following observation of a model's behaviour, the trustor updates its beliefs about the model's reliability and competence under uncertainty and ambiguity towards achieving the objectives of the study. The beliefs are moderated via multiple factors, including the importance of the situation to model user, the utility gained (or loss experienced), and the tolerance level for using the model. The acceptability threshold indicates how much one needs to trust a model to decide to use it.

The perceived risk in using the model and competence of the model together provide a basis for deciding the acceptability threshold of the model. Increased competence of the model lowers the acceptability threshold, making it easier to trust the model. On the other hand, increased perceived risk would increase the threshold. The importance of the modelling situation can moderate these two factors. In a given interaction step with the model in a specific context, the trust measure can be updated by an amount that is a function of the acceptability threshold, the situational trust, and the affective state (e.g., regret, forgiveness) because trust is only required if an adverse outcome would make us regret our decision (Luhmann, 1979). The situational trust is determined by the utility gained (i.e., the value of solving the specific problem), its importance, and the general trust attitude (disposition) towards the model. If the situational trust is higher than the acceptability threshold in a specific context, the model can be considered as sufficiently trustworthy for use in that situation.

6. Measurement of trust

The level of detail in conceptualising the outcome of trust evaluation can vary depending on the problem domain, the characteristics of the users of the model, and the characteristics of the model. As shown in Figure 5, based on the characteristics of model users, three prototype evaluation models with an increasing level of specificity can be considered: binary evaluation, quantised/discrete evaluation, and continuous/spectral evaluation.

Each evaluation type depends on the model user's level of concern with the problem, cognitive ability, expertise or experience in the problem domain, and the availability of reference frames that can be used to compare the model's utility to desirable outcomes. Table 2 depicts how these criteria relate to the three types of evaluation models.

6.1. Binary evaluation

At the highest level of abstraction, users can categorise a model as either trusted or distrusted without any middle ground. Such a binary decision is reasonable when model users have the following characteristics: (1) low-level concern with the problem (2) limited ability to process information due to comprehension skills or contextual ambiguity and uncertainty, (3) low-level familiarity, expertise, and experience in the domain of the problem, and (4) lack of standards or reference frames for comparison.

6.2. Quantised evaluation

Quantisation of the trust measurement, scaled into a limited range of discrete values, enables a more precise evaluation of trust. By setting thresholds over the continuous measurement range supports the classification of the degree of trust and hence increases the level of resolution for decision-support. As shown in Figure 5, the acceptability threshold within the favourable region

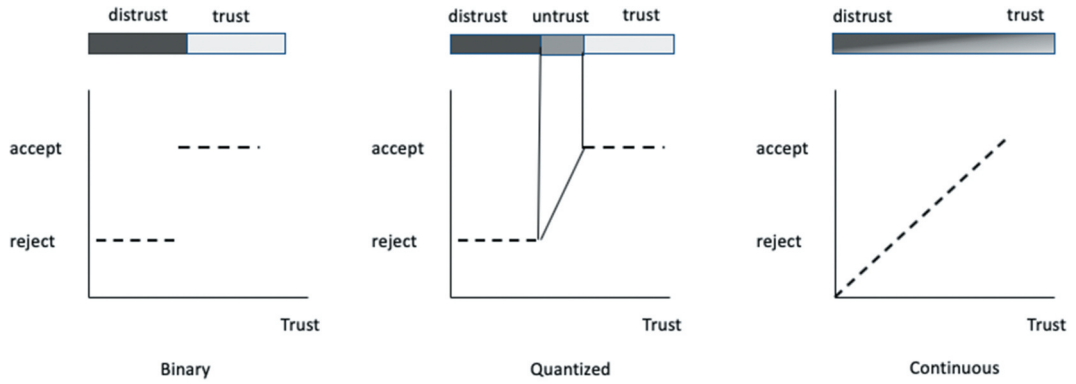


Figure 5. Trust measurement.

Table 2. Types of trust evaluation.

Model of Trust (Credibility) Evaluation	Based on the criteria: concern with the problem	ability to process information	domain familiarity/ experience/ expertise	availability of standards/ reference models
Binary	low	limited	low	lacks standards (none)
Quantised	moderate	moderate	partial	some
Spectral	high	significant	high	substantial

distinguishes the state of trust from untrust, which characterises the situation when the model user cannot have sufficient confidence but lacks evidence for either trusting or distrusting the model. On the other hand, a threshold in the unfavourable region distinguishes undistrust from distrust to cover the cases when the model user lacks trust. Discrete categories of trust can be refined with fuzzy values (e.g., low, medium, high) to indicate the incremental strength of trust within each category. Discrete evaluation is appropriate when model users have the following characteristics: (1) moderate level of interest in the problem, (2) moderate cognitive ability to process information due to comprehension skills or contextual ambiguity and uncertainty, (3) partial familiarity, expertise, and experience in the domain of the problem, and (4) partial availability of standards or reference frames for comparison.

6.3. Continuous (Spectral) evaluation

Measuring trust over a continuous range without quantisation gives model users the discretion for deciding on the degree and strength of trust. The spectral strategy is appropriate when model users have (1) high-level concern with the problem, (2) significant cognitive ability to process information in the presence of favourable contextual factors, (3) high familiarity, expertise, and experience in the domain of the problem, (4) significant availability of standards or reference frames for comparison.

7. Considerations for appropriate trust in models

Inappropriate reliance on a model, resulting in misuse or disuse, is often due to mismatch between perceived trust and model capabilities. When trust exceeds the capabilities of a model due to poor calibration, overtrust gives rise to invalid decisions and conclusions based on the outcome of the model. On the other hand, with inappropriate distrust, trust falls short of the capabilities of the model and hence results in loss of opportunity. In this section, we present a set of guidelines for engineering trustable models.

7.1. Design for appropriate trust

Models need to be transparent to be effectively communicated to users to facilitate matching the capabilities of a model with the desired level of trust. This requirement can be achieved by providing modular and explainable algorithms that reveal their semantics and pragmatics more clearly. The following guidelines and considerations aim to support the design and evaluation of trustable models.

- Instead of designing models for maximised trust, focus on design for an appropriate context-sensitive trust that mitigates overtrust as well as distrust.
- Provide information about and access to prior use of the model and the explanatory models of the outcomes of simulation experiments.
- Reveal the causal mechanisms and processes that generate a model's behaviour so that observed behaviour can be traced to the underlying assumptions and intermediate results.
- Make the purpose of the model and the research inquiry, which the model is designed for explicit. Both the design rationale and the model's purpose need to be easily related to the objectives of the simulation experiment.
- Train model users regarding the acceptable uses of the model and its limitations.

These considerations are predicated on the trade-offs between trustworthy models and trustable models. While a trustworthy model can be dependable and reliable, its design may be overly complicated, difficult to understand, and hence less trustable. To the extent that a model's adoption requires its ease of use and maintenance, there may be circumstances that favour simplified, less capable models that are easier to calibrate.

7.2. Relating context to model capability

The relationship between a model and its referent is meaningful in the context of their use by a cognitive agent. That is, a model is not simply a representation of the referent. It is used by a cognitive agent (e.g., scientist, engineer, artificial agent) to make sense of the system under study. Therefore, the role of the cognitive agent needs to be explicit. Concerning the context, specific considerations for trustable models include the following:

- Specify the context and objectives explicitly in a way that relates the elements of a model to its context. The association between the elements of the model and the context is expected to be many-to-many. A model feature may be necessary for multiple research inquiries. Similarly, a contextual element may require the provision of multiple features configured in a meaningful manner.
- Describe the past performance of the model concerning contextual scenarios. Classifying a model's

outcomes over the context information facilitates reasoning about its robustness, which is an important trust metric under uncertainty.

• Evaluate the impact and influence of context on the pertinence of trust. By determining the relation between context and trust, model users can reason about the potential mismatch between the model's capabilities and trust in a context-sensitive manner and use the model selectively.

• Characterise the trustworthiness of a model in terms of how well model capabilities match the extent of trust. Having too many capabilities, each with low trust, is as undesirable as is a non-uniform distribution of trust across capabilities. Trust should be broad and uniform across capabilities while being cognisant of the significance of context on trust. Moreover, the specificity of trust defined by the association of trust with specific elements of a model is critical when the performance is context-sensitive. A specific component of a model may not exhibit an acceptable level of reliability across all contextual situations.

7.3. The influence of domain, field, and individual interactions on trust

Model development does not take place in isolation. It is a creative endeavour that is influenced by three major components of a technical discipline. Each discipline has established norms, methods, research problems, evaluation criteria, and domain knowledge that together provide a frame of reference for the relevant constructs of a model. Contributions made by individuals are evaluated by the members of the domain (i.e., field), and those contributions accepted by the field are then included in the domain. Individuals are acculturated and trained in the norms and practices of a domain to develop solutions (e.g., models) that are sufficiently consistent with the domain knowledge. The field uses the normative evaluation criteria and serves as a gatekeeper by deciding on the trustworthiness of knowledge for inclusion in the domain. Awareness of these interactions and related factors are integral in developing appropriate trust. Considerations in the design and evaluation of models from the perspective of organisational interaction include the following:

• Recognise the characteristics of the context of both the domain and its field along with their direct and indirect influence on the evaluation of models.

• Develop awareness about individual and domain-specific cultural differences in using criteria for trust assessment. An operations research group that favours empirical accuracy and similarity concerning a referent system is expected to have different criteria for trust than a research group in life sciences exploring possible causal mechanistic explanations of a phenomenon.

• Cultural differences can result in different expectations from a model. For instance, a research group that views trust as forward-looking considers models that can apply to future problems as more trustable compared to a model that fits data better than the alternatives.

• Recognise the importance of reputation and relational factors, including but not limited to coherence with reference models, in instilling confidence about the utility of a model.

8. Conclusions

Simulation models have become integral components of research in the overall science and engineering enterprise. The increasing use and the broad range of purposes of computational models necessitate revisiting the notion of model credibility that has long been considered as a measure of similarity and accuracy. Besides prediction, models are being used to explore possibilities, discover explanations for scientific phenomena, conduct training and tutoring in virtual environments, and control physical systems. In practice, model credibility is viewed not only as a function of its features but also from the perspective of the cognitive interests of model users in a specific context of inquiry.

Factoring in the cognitive perspective and the contextual interpretation of the multi-faceted characterisation of model credibility can be facilitated by associating credibility with trust. Despite the co-occurrence of credibility and trust as keywords in an increasing number of publications, especially in computational modelling for social and behavioural sciences, the current view of credibility within the M&S community is still centred around the notion of accuracy and empirical fitness. To mitigate this limitation, we characterised model credibility in terms of the attributes, properties, and dynamics of trust. Three dimensions of trust and associated concepts are used to provide a foundation for trust-driven assessment of credibility. Fundamental processes in the formation and evolution of trust are highlighted to facilitate maintaining run-time trust models that can serve as cognitive agents to establish appropriate, calibrated trust. Based on the characteristics of the model users, alternative evaluation models are recommended to measure and report trust metrics. We conclude the paper with recommendations for establishing appropriate trust in simulation models. These considerations highlight design principles, provide strategies for relating contextual information to model capabilities for context-sensitive trust measurement and emphasise the significance of factoring into the process of the influence of the organisational culture and its community of practice.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported by the National Science Foundation [NSF-IIS-1910794].

References

Balci, O. (1986). Credibility assessment of simulation results. In *Proceedings of the 18th conference on winter simulation* (pp. 38–44). Washington DC, USA.

Balci, O. (2015). Quality indicators throughout the modeling and simulation life cycle. In Levent Yilmaz (Ed.), *Concepts and methodologies for modeling and simulation* (pp. 199–215). Springer.

Casti, J. L. (1997). Can you trust it? *Complexity*, *2*(5), 8–11. https://doi.org/10.1002/(SICI)1099-0526(199705/06)2:5<8::AID-CPLX2>3.0.CO;2-3

Chen, C. (2006). Citespace ii: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, *57*(3), 359–377. https://doi.org/10.1002/asi.20317

Davis, P. K., O'Mahony, A., Gulden, T. R., Osoba, O. A., & Sieck, K. (2018). *Priority challenges for social and behavioral research and its modeling*. RAND Corporation Santa Monica, CA.

Dempster, A. P. (2008). Upper and lower probabilities induced by a multivalued mapping. In Roland R. Yager & Liping Liu (Eds.), *Classic works of the dempster-shafer theory of belief functions* (pp. 57–72). Springer.

Ebert, T. A. (2009). Facets of trust in relationships–a literature synthesis of highly ranked trust articles. *Journal of Business Market Management*, *3*(1), 65–84. https://doi.org/10.1007/s12087-008-0034-9

Falcone, R., & Castelfranchi, C. (2001). Social trust: A cognitive approach. In Castelfranchi, Cristiano, Yao-Hua Tan (Eds.), *Trust and deception in virtual societies* (pp. 55–90). Springer.

Fogg, B., & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 80–87). Pittsburgh, Pennsylvania.

Fulmer, C. A., & Gelfand, M. J. (2012). At what level (and in whom) we trust: Trust across multiple organizational levels. *Journal of Management*, *38*(4), 1167–1230. https://doi.org/10.1177/0149206312439327

Gambetta, D. (2000). Can we trust trust. In Gambetta, D (ed.) *Trust: Making and Breaking Cooperative Relations, electronic edition*, (chapter 13, pp. 213-237). Department of Sociology, University of Oxford. https://www.csee.umbc.edu/~msmith27/readings/public/gambetta-2000a.pdf

Gelfert, A. (2016). *How to do science with models: A philosophical primer*. Springer.

Gelfert, A. (2019). Assessing the credibility of conceptual models. In Claus Beisbart and Nicole J. Saam (Eds.), *Computer simulation validation* (pp. 249–269). Springer.

Hardwig, J. (1991). The role of trust in knowledge. *The Journal of Philosophy*, *88*(12), 693–708. https://doi.org/10.2307/2027007

Harper, A., Mustafee, N., & Yearworth, M. (2020). Facets of trust in simulation studies. *European Journal of Operational Research*. https://doi.org/10.1016/j.ejor.2020.06.043

Janani, V., & Manikandan, M. (2018). Mobility aware clustering scheme with bayesian- evidence trust management for public key infrastructure in ad hoc networks. *Wireless Personal Communications*, *99*(1), 371–401. https://doi.org/10.1007/s11277-017-5107-1

Johnson, J. (2000). The "can you trust it?" problem of simulation science in the design of socio-technical systems. *Complexity*, *6*(2), 34–40. https://doi.org/10.1002/cplx.1017

Knutti, R. (2008). Should we believe model predictions of future climate change? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1885), 4647–4664. https://doi.org/10.1098/rsta.2008.0169

Knuuttila, T. (2005). Models, representation, and mediation. *Philosophy of Science*, *72*(5), 1260–1271. https://doi.org/10.1086/508124

Larsen, L., Thomas, C., Eppinga, M., & Coulthard, T. (2014). Exploratory modeling: Extracting causality from complexity. *Eos, Transactions American Geophysical Union*, *95*(32), 285–286. https://doi.org/10.1002/2014EO320001

Lee, E. A. (2008). Cyber physical systems: Design challenges. In *2008 11th ieee international symposium on object and component-oriented real-time distributed computing (isorc)* (pp. 363–369). Orlando, Florida.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Lerch, F. J., & Prictula, M. J. (1989). How do we trust machine advice? In *Proceedings of the third international conference on human-computer interaction on designing and using human-computer interfaces and knowledge based systems* (2nd ed., pp. 410–419). Elsevier.

Levins, R. (1966). The strategy of model building in population biology. *American Scientist*, *54*(4), 421–431. https://doi.org/10.2307/27836590

Luhmann, N. (1979). *Trust and power*. John Wiley & Sons.

Marsh, S., & Briggs, P. (2009). Examining trust, forgiveness and regret as computational concepts. In Golbeck, Jennifer (Ed.), *Computing with social trust* (pp. 9–43). Springer.

Marsh, S. P. (1994). *Formalising trust as a computational concept*. Technical Report - Ph.D. Dissertation, Department of Computing Science and Mathematics, University Stirling. https://www.nr.no/~abie/Papers/TR133.pdf

Merriam-Webster. (2019). *Merriam-webster*. Springfield, Massachusettas: Merriam-Webster, Incorporated. Online at https://www.merriam-webster.com/.

Nilsson, N. J. (2014). *Understanding beliefs*. MIT Press.

Onggo, S., Yilmaz, L., Klugl, F., Terana, T., & Macal, C. M. (2019). Credible agent- based simulation – An illusion or only a step away? In *2019 Winter Simulation Conference (WSC) 2019 Dec 8* (pp. 273–284). IEEE.

Parker, W. S. (2009). Confirmation and adequacy-for-purpose in climate modelling. *Proceedings of the Aristotelian Society, Supplementary Volumes, 83*, 233–249. Retrieved September 12, 2020, from http://www.jstor.org/stable/2061913

Peters, H. P., & van Raan, A. F. (1993). Co-word-based science maps of chemical engineering. part i: Representations by direct multidimensional scaling. *Research Policy*, *22*(1), 23–45. https://doi.org/10.1016/0048-7333(93)90031-C

Radhakrishnan, S., Erbis, S., Isaacs, J. A., & Kamarthi, S. (2017). Novel keyword co-occurrence network-based methods to foster systematic reviews of scientific literature. *PloS One*, *12*(3), e0172778. https://doi.org/10.1371/journal.pone.0172778

Rossiter, J. A. (2017). *Model-based predictive control: A practical approach*. CRC press.

Rotter, J. B. (1980). Interpersonal trust, trustworthiness, and gullibility. *American Psychologist*, *35*(1), 1. https://doi.org/10.1037/0003-066X.35.1.1

Sargent, R. G. (1983). *Validating simulation models* (Tech. Rep.). Institute of Electrical and Electronics Engineers (IEEE).

Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, *7*(1), 12–24. https://doi.org/10.1057/jos.2012.20

Sargent, R. G., & Balci, O. (2017). History of verification and validation of simulation models. In *Proceedings of the 2017 winter simulation conference* (p. 17). Las Vegas, NV, USA.

Self, C. C. (2014). Credibility. In Don W. Stacks and Michael B. Salwen (Eds.), *An integrated approach to communication theory and research* (pp. 449–470). Routledge.

Thagard, P. (2002). *Coherence in thought and action*. MIT press.

Tolk, A., Barros, F., D'Ambrogio, A., Rajhans, A., Mosterman, P. J., Shetty, S. S., … Yilmaz, L. (2018). Hybrid simulation for cyber physical systems: A panel on where are we going regarding complexity, intelligence, and adaptability of cps using simulation. In *Proceedings of the symposium on modeling and simulation of complexity in intelligent, adaptive and autonomous systems* (p. 3). Baltimore, Maryland.

Van Eck, N. J., & Waltman, L. (2010). Software survey: Vosviewer, a computer program for bibliometric mapping. *scientometrics*, *84*(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3

Wang, S. W., Ngamsiriudom, W., & Hsieh, C.-H. (2015). Trust disposition, trust antecedents, trust, and behavioral intention. *The Service Industries Journal*, *35*(10), 555–572. https://doi.org/10.1080/02642069.2015.1047827

Weisberg, M. (2012). *Simulation and similarity: Using models to understand the world*. Oxford University Press.

Yilmaz, L. (2004). On the need for contextualized introspective models to improve reuse and composability of defense simulations. *The Journal of Defense Modeling and Simulation*, *1*(3), 141–151. https://doi.org/10.1177/875647930400100302

Yilmaz, L. (2006). Validation and verification of social processes within agent-based computational organization models. *Computational & Mathematical Organization Theory*, *12*(4), 283–312. https://doi.org/10.1007/s10588-006-8873-y

Young, P. (1983). The validity and credibility of models for badly defined systems. In Beck, M.B., Straten, G. van (Eds.), *Uncertainty and forecasting of water quality* (pp. 69–98). Springer.