Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

EXPLAINABILITY IN DIGITAL TWINS: OVERVIEW AND CHALLENGES

Meryem Mahmoud[1] and Sanja Lazarova-Molnar[1, 2]

[1]Institute AIFB, Karlsruhe Institute of Technology, Karlsruhe, GERMANY
[2]The Maersk Mc-Kinney Moller Institute, University of Southern Denmark, Odense, DENMARK

ABSTRACT

Digital Twins are increasingly being adopted across industries to support decision-making, optimization, and real-time monitoring. As these systems and, correspondingly, the underlying models of their corresponding Digital Twins, grow in complexity, there is a need to enhance explainability at several points in the Digital Twins. This is especially true for safety-critical systems and applications that require Human-in-the-Loop interactions. Ensuring explainability in both the underlying simulation models and the related decision-support mechanisms is key to trust, adoption, and informed decision-making. While explainability has been extensively explored in the context of machine learning models, its role in simulation-based Digital Twins remains less examined. In this paper, we review the current state of the art on explainability in simulation-based Digital Twins, highlighting key challenges, existing approaches, and open research questions. Our goal is to establish a foundation for future research and development, enabling more transparent, trustworthy, and effective Digital Twins.

1 INTRODUCTION

Digital Twins (DTs) have emerged as transformative tools that enable dynamic digital replicas of physical systems, enabling real-time monitoring, simulation, and decision-making across various industries (Boschert and Rosen 2016). By mirroring behaviors of their physical counterparts, DTs use simulation techniques —such as Discrete Event Simulation, System Dynamics, and Agent-Based Modeling—with stochastic elements and conditional logic (Law 2015).

These simulation techniques are theoretically transparent at the component level, with rules that are explicitly defined to enable traceability (Law 2015). However, when integrated into a large-scale DT framework, their emergent complexity makes it hard to track how specific inputs lead to certain outputs, even to domain experts (Riis et al. 2022). Simulation models can become overly complex rapidly due to a large number of variables, potential interactions between these variables, and possible non-linear effects (Nigel and Klaus 2005). This complexity makes DTs hard to interpret, giving them the appearance of black-box models (Lorscheid, Heine, and Meyer 2012). Particularly in safety-critical domains such as nuclear energy, aerospace, and autonomous vehicles, decisions hinge on understanding how components collectively generate outcomes (Koopman and Wagner 2017). For instance, optimization strategies lacking interpretability may produce counterintuitive recommendations (e.g., prioritizing cost savings over safety), eroding trust even when models are theoretically traceable (Yang et al. 2025).

The ambiguity in how inputs propagate through interconnected subsystems is aggravated by three key factors. First, the inherent complexity of modern physical systems, which DTs are designed to replicate. Second, the heterogeneity of data sources, formats, and communication protocols across subsystems (e.g., sensors, Internet of Things (IoT) devices, multimodal data streams, simulations), requiring an underlying data infrastructure capable of integrating diverse inputs and enabling seamless communication (Friederich et al. 2022). Third, inefficient human-machine cooperation, as simulation models and their analyses are

979-8-3315-8726-0/25/$31.00 ©2025 IEEE

often not described exhaustively due to publication limitations or to avoid overwhelming the audience (Axelrod 1997).

Therefore, the need for explainability has become paramount. For example, in domains such as healthcare, and manufacturing, decisions must be traceable, ethical, and safe to build trust, support effective decision-making, and drive widespread adoption (Giabbanelli 2024; Zhang et al. 2024a). Stakeholders require not just technical accuracy but also human-understandable narratives that validate how inputs map to outcomes.

The goal of this paper is to provide an overview of the importance of explainability in DTs, particularly focused on DTs that feature simulation models as underlying models. The paper is organized as follows. Section 2 provides a background on the concepts of Digital Twins and Explainability, followed by a deeper analysis of Explainability within Digital Twins. In Section 3, we review the limited existing approaches that attempt to integrate explainability in simulation-driven DTs. In Section 4, we identify the challenges and open questions related to this domain. Finally, Section 5 provides a summary and future research directions.

2 BACKGROUND

As DTs continue to evolve, their reliance on real-time data and simulation models has become increasingly important. In this section, we introduce the core components of DTs, emphasizing their data-driven nature and simulation-based frameworks. In addition, we explore the concept of explainability and its importance, setting the stage for its integration in simulation-based DTs.

2.1 Digital Twins

DTs are an evolution of traditional simulation modeling, as both focus on understanding, monitoring, and analyzing physical systems (Law 2015). While traditional simulations are static and scenario-bound, DTs evolve continuously through bidirectional data exchange, allowing for predictive maintenance, adaptive decision-making, and continuous system improvement. Friederich et al. (2022) define Digital Twins as comprising of three core components:

- The real-world entity, which can range from one single process to an entire operation.
- A data-driven simulation model, which includes algorithms for modeling as well as connectivity components for real-time data exchange.
- The data collected from the physical system ensures the accuracy and effectiveness of the DT.

Building on this foundation, Figure 1 presents a data-driven DT framework that reflects an iterative process starting with data generation. The data flows through interconnected stages: Collection, Validation, Knowledge Extraction, Model Development, and Model Validation. Notably, "Analysis" is positioned to reflect its ongoing interaction with the simulation model—both informing and being informed by it—underscoring the cyclical nature of insight generation. This continuous synchronization between the DT and the real-world entity ensures high fidelity and supports robust system optimization.

Central to the data-driven DT framework is the data-driven simulation model, which employs simulation techniques such as Discrete Event Simulation (DES), System Dynamics (SD), and Agent-Based Modeling (ABM) to replicate system behavior (Maidstone 2012; Law 2015). These models dynamically update via sensor data, IoT connectivity, or operational databases, forming the backbone of simulation-based DTs. In automotive DTs validate radar sensors for automated driving by simulating real test drives in virtual environments and comparing sensor outputs using statistical metrics (Magosi et al. 2022). They also simulate production line reconfigurations to optimize manufacturing efficiency (Yang et al. 2022). The energy sector leverages DTs to model grid behavior for renewable integration and real-time load optimization (Ghenai et al. 2022), while simulating photovoltaic systems and fuel cell plants for fault diagnosis and operational safety. In healthcare, patient-specific DTs simulate treatment responses to tailor personalized therapies (Kamel Boulos and Zhang 2021). For smart cities, GIS-integrated DTs support traffic management simulations to reduce congestion, and traffic trace data is used to optimize urban-scale energy efficiency (Abdeen and Sepasgozar 2021).

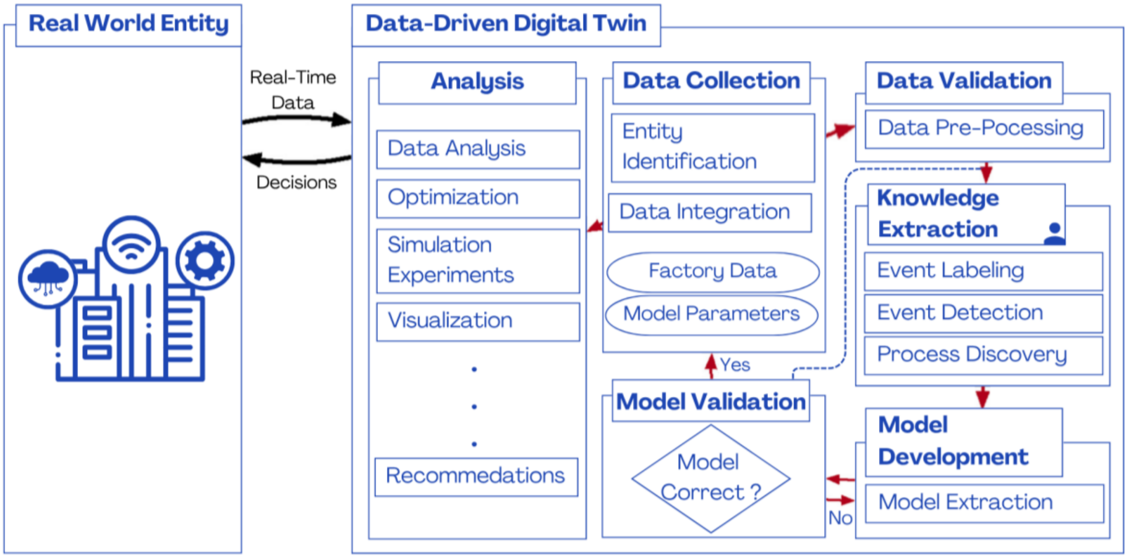

Figure 1 : Framework for Data-Driven Digital Twins (Lazarova-Molnar 2024; Friederich et al. 2022)

2.2 Explainability

Explainability refers to the extent to which a model's internal processes and outputs can be understood by its human users (Miller 2019). In recent years, the growing emphasis on explainability has stemmed from ethical concerns, regulatory requirements (Goodman and Flaxman 2017) and the need for trust, validation, and actionable insights (Miller 2019).

The rise of complex black-box models, such as neural networks, has intensified this need. These black-box models rely on non-linear transformations, hidden layers parameters, and uninterpretable weights, making their logic opaque even to developers (Hamm et al. 2023). Consequently, this sparked the development of explainable AI (XAI) methods, such as post hoc interpretability techniques— including Local Interpretable Model-agnostic Explanations (LIME) and Shapley Additive Explanations (SHAP) (Hamm et al. 2023; Lundberg and Lee 2017)— and the use of inherently transparent models (e.g., linear regression models, rule-based systems, or decision trees) that are fully understandable by design (Adadi and Berrada 2018).

As discussed in depth in Vilone and Longo's work (2021), effective explanations must meet several key requirements to ensure clarity and usefulness. Some of the essential attributes include completeness, contrastivity, user-centricity, and actionability. Completeness ensures that the explanation faithfully represents the model's reasoning without oversimplification. Contrastivity highlights why a specific prediction was made over alternatives, helping users understand the decision-making process. User-centricity tailors explanations to the audience's expertise. Finally, actionability enables users to correct model errors or adjust input based on the explanation provided.

The modalities of explanations vary depending on the audience and context. Different users require distinct types of explanations depending on their role and needs. Glomsrud et al. (2019) identified four types of explanations demanded by the different stakeholders, which we generalize as follows. Developer explanations that provide in-depth technical details for researchers and developers who need to understand, verify, or refine the system. These explanations are often complex and require deep technical expertise. Assurance explanations which focus on demonstrating the system's reliability and compliance with specific requirements, helping stakeholders validate its performance and trustworthiness. End-user explanations are designed for individuals who interact with the system directly, offering real-time reasoning in a clear and intuitive way without requiring knowledge of the system's internal workings. Finally, external explanations are aimed at those affected by a system's decisions but not directly involved in its operation, presenting the system's rationale in an accessible and interpretable manner.

Common forms of explanations include textual, visual, auditory, and multimodal explanations. Each of these modalities serves different purposes and is suited to different contexts. Textual explanations provide natural language rationales, often generated via rule-based systems (Rodis et al. 2024) or neural text generation approaches such as Large Language Models (LLMs) (Zhang et al. 2024b; Yang et al. 2025). Visual explanations use graphical elements, such as saliency maps and heat maps, to illustrate the reasoning behind decisions (Vilone and Longo 2021). Auditory explanations convey information using sound and are emerging in fields such as speech recognition (Wu, Bell, and Rajan 2023). Although less common, these explanations can enhance accessibility, particularly for visually impaired users. Multimodal explanations combine two or more modalities to enhance explainability. These explanations can be synchronous or asynchronous, integrating textual, visual, and auditory elements for a more comprehensive understanding (Rodis et al. 2024).

Having specified the concept of explainability, it is important to note that while some DT models are based on machine learning techniques, often seen as black-box models, we are specifically interested in the explainability of *simulation-based DTs*. In Section 3, we discuss explainability in simulation models and DTs and review the existing literature on this subject.

3 OVERVIEW OF EXPLAINABILITY IN DIGITAL TWINS

Explainability in DTs is essential for ensuring transparency and trust in complex simulation models, as well as the resulting decisions. In this section, we explore how explainability can be integrated into the different components of DTs. We review existing explainability approaches and frameworks to highlight progress and identify research gaps.

3.1 Explainability in Digital Twins

Explainability in DTs ensures that users can understand how a DT functions, with visualization being one of the key tools in making the outputs comprehensible (Ali et al. 2024). This is challenged by both the complexity of the physical systems they mirror and the technical limitations of the modeling process.

As modern systems grow increasingly complex, they incorporate multiple interconnected subsystems that generate vast volumes of heterogeneous data. While these subsystems contribute to the overall functionality, they also introduce new dependencies complicating the transparency and explainability of DTs. For instance, to enhance reliability and ensure fault tolerance, less reliable components are often supplemented with redundancies, enabling systems to maintain operation even when individual components fail (Sun et al. 2025). Similarly, dense sensor networks facilitate continuous data collection, providing real-time insights while significantly increasing storage and data management demands (Correia, Abel, and Becker 2023). This redundancy hinders traceability, making it difficult to find root causes for anomalies.

When it comes to the modeling process, even though simulation models (e.g., DES, DS, ABM) are perceived as inherently explainable, their transparency fades as models scale. Creating accurate simulation models is time-consuming and labor-intensive (Yang et al. 2025), often integrating rules and conditions alongside stochastic elements. However, the combination of randomness and multiple conditional possibilities can make it hard to fully understand why the simulation behaves in certain ways across different scenarios (Grigoryan 2024). Consequently, the modeling process itself creates significant barriers to both the prediction accuracy and the interpretability of the outcomes, as tracing the logic behind specific results in the simulation is impractical without a thorough understanding of the underlying model design.

Understanding the reasoning behind a simulation model's output and recommendations often requires extensive domain-specific expertise, as its outputs typically include complex formats such as measurements and sensor readings that are difficult to parse (Zhang et al. 2024b). In some cases, simulation-derived system recommendations may even be counterintuitive, emphasizing the need for the rationale behind them to facilitate further decision-making and actionability (Zhang et al. 2024a). Therefore, in human-in-the-loop applications, DTs must provide intuitive and user-friendly explanations, considering the expertise of the stakeholders (Yang et al. 2025; Zhang et al. 2024a).

In the following, we view explainability through the lens of our earlier-introduced framework that we illustrate in Figure 1. Explainability can be used to enhance the different elements as follows:

- Data Validation, where explainability can support the process of validating data as a critical component of data-driven simulation and data-driven DTs,
- Knowledge Extraction, where explainability can support the process of knowledge extraction that precedes the model extraction and development processes,
- Model Development, where explainability can support the model extraction processes by providing insights into how and why certain model elements have been extracted,
- Model Validation, where explainability can support face validity, as well as credibility of the quantitative validation processes,
- Decisions, where explainability can support the simulation-based (or optimization-based) resulting decision support.

We use these elements as a basis to study available literature on explainability in simulation and DTs, detailed in the following subsection.

3.2 Existing Approaches for Explainability in DTs and Simulation

To identify relevant work, we surveyed existing scientific literature across IEEE Xplore, ACM, ScienceDirect, SpringerLink, and Google Scholar using the following keywords: “explainability,” “simulation-based Digital Twins,” “interpretability,” as well as more specific phrases such as “interpretable Digital Twin”, “transparent Digital Twin” and “white-box Digital Twin”. We identified a research gap in the existing literature as limited work has been conducted on this topic, particularly with respect to DTs. Consequently, we expanded our search to include general simulation explainability research as well as human-in-the-loop contexts. Table 1 synthesizes these findings, focusing on approaches that integrate explainability into both DTs and simulations, while highlighting their explainability goal, explainability approach, Digital Twin element, and application domain. The 'Explainability Goal' column in Table 1 reflects the primary objectives as articulated by the authors, extracted through careful reading and interpretation of the selected studies to ensure accurate representation of their stated or implied intentions. These studies highlight the diverse strategies researchers have employed to enhance transparency in DTs. We note that all approaches focus on isolated aspects of DTs rather than addressing the full lifecycle of these systems.

3.3 Discussion

Table 1 outlines four main approach categories: LLM-driven explanations, post-hoc interpretability methods, collaborative Human-in-the-Loop frameworks, and self-explainability frameworks.

3.3.1 On Large Language Model-driven Explanations

The integration of LLMs represents a significant advancement in explainability in DTs (Blasek et al. 2023; Yang et al. 2025) with their accessibility increasing due to the rise of tools such as ChatGPT. LLM-driven explanations use models trained on extensive corpora to generate context-aware explanations through techniques such as retrieval-augmented generation and prompt engineering (Gao et al. 2023; Liu et al. 2023). Leveraging LLM-driven explanations, significantly lowers the expertise barrier, offering natural language explanations for DTs.

Table 1: Existing Approaches for Explainability in DTs and Simulation.

Application Domain	Digital Twin Element	Explainability Goal	Summary of the Explainability Approach	Reference
Smart Manufacturing	Model Development, Model Validation	Stakeholders' collaboration framework during modeling phase, to enhance their trust in the DT	- Mathematical formulation of modeling processes - Three types of explanations: model-based, scenario-based, and goal-oriented - Development of explainability scores based on option differences and performance impacts	Wang et al. (2021)
General	Knowledge Extraction, Model Validation, Decisions	Integrating LLMs into DT framework to provide explicable decision-making	- Integrating LLMs throughout the DT lifecycle - Enhancing explainability by using LLM-enhanced data analysis and strategy explanation	Yang et al. (2025)
Smart Agriculture	Decisions	Providing a natural language explanation for decisions made by Dynamic data-driven DTs	- Integrating LLM explainability in the DDT depending on the decision maker (DDT, Human, LLM) - Generating textual explanations for autonomous decisions through retrieval-augmented generation	Zhang et al. (2024b)
Healthcare	Decisions	Generating narrative explanations using LLMs for translating graphical representation of (scaled up) agent-based simulation model into accessible formats	- Agent-based models for simulating population-scale interventions - Translating simulation model visualizations (e.g., node and link diagrams) into textual reports (graph-to-text) for clinicians, policymakers and community members	Giabbanelli (2024)
Game Theory	Model Development, Decisions	Using feature importance to explain how different input parameters affect the results of a simulation model	- Using post-hoc Shapley value calculation and Nucleolus-based methods to determine feature importance	Grigoryan (2024)
Air Traffic Management	Model Development, Decisions	Combines simulation meta-modeling with SHAP to provide functional approximations and quantify the effect each input variable has on the output	- Simulation metamodeling using XGBoost - Integrating SHAP values for feature importance analysis	Riis et al. (2022)
Smart Manufacturing	Data Collection, Model Development, Model Validation	Explaining autonomous decisions made by the DT to human operators using interpretable machine learning techniques.	- Architecture with three controllers implementing interpretable machine learning (K-Nearest Neighbors, Support Vector Machines, decision tree etc.) - Adaptive selection algorithms for dynamic model switching.	Zhang et al.(2024a)
Automotive	Model Development, Model Validation, Decisions	Enhance self-explainability of DTs through hierarchical, model-driven explanations	- Use of the MAB-EX framework (Monitor, Analyze, Build, Explain) to enable DTs to explain their decisions - Deriving explanations from system, process, and reasoning models - Tailoring explanations to different stakeholders.	Michael et al. (2024)

The work of Yang et al. (2025) present a comprehensive technical framework illustrating how LLMs can be integrated with simulation models throughout the DT lifecycle. Explainability is particularly relevant in enhancing data analysis, improving the interpretation of complex patterns and relationships, and aiding strategy explanation. In smart agriculture, LLMs empower farmers by generating explanations for autonomous drone monitoring decisions (Zhang et al. 2024b). Similarly, Giabbanelli's (2024) work in healthcare shows how LLMs can be used to generate narrative explanations for decisions based on agent-based simulation models of mental health interventions, aiding policymakers in assessing the fairness and effectiveness of various strategies.

However, challenges remain in ensuring the reliability and accuracy of LLM-generated explanations, particularly in critical decision-making contexts. Computational demands (Yang et al. 2025) and the potential for hallucinations due to inadequate domain adaptation (Huang et al. 2025) pose significant challenges that can undermine user trust.

3.3.2 On Post-hoc Interpretability Methods

Post-hoc explanations are techniques used to examine and understand a model's decision-making process after it generates predictions, providing insight into how its outputs were determined (Retzlaff et al. 2024). A widely used post-hoc approach involves assessing feature importance to understand the impact of individual variables on the model's output (Adadi and Berrada 2018). Among these techniques, SHAP is particularly notable. Originally introduced by Shapley (1953) for game theory; it was later adapted by Lundberg and Lee (2017) to enhance interpretability in machine learning models.

Riis et al. (2022) and Grigoryan (2024) utilized Shapley values to quantify the contribution of individual features to model predictions, providing explanations that enhance transparency and trustworthiness. Riis et al. applied SHAP to Air Traffic Management (ATM) simulations to interpret how input parameters (e.g., fuel prices, planning horizons) affect performance metrics such as passenger delays. Grigoryan extended SHAP to agent-based predator-prey models, demonstrating how features such as predator reproduction rates and resource availability drive emergent behaviors.

The technical nature of these explanations may be difficult for non-experts to interpret without additional visualization or simplification, particularly in stochastic systems that heavily rely on mathematical annotations (Grigoryan 2024). However, a key limitation of SHAP is that these explanations focus solely on input relationships rather than directly mapping to the simulator's actual output behavior. Additionally, the calculation of Shapley values involves evaluating the model's performance across all subsets of features, ensuring a fair and comprehensive assessment of each feature's importance. This process comes at a high computational cost, which scales exponentially with input dimensionality, necessitating strategies such as active learning to reduce training data requirements (Riis et al. 2022).

3.3.3 On Collaborative Human-in-the-Loop Frameworks

Collaborative frameworks represent a promising direction for enhancing explainability in DTs by integrating human expertise with automated systems. Effective decision-making relies on the synergy of computational analysis and human intervention. As noted by Wang et al. (2021), the integration of stakeholders in the modeling process is essential for the successful implementation of DTs. The framework of Wang et al. automates the generation of three types of explanations: (1) model-based explanations, which clarify how different model configurations impact performance metrics, providing a comparative analysis of model options; (2) scenario-based explanations, which interpret how model performance varies across different business scenarios, offering insights into model robustness and adaptability; and (3) goal-oriented explanations, which guide how models can be modified to achieve specific performance objectives, often formulated as optimization tasks. To quantify the value of explanations, informativeness of explanations is used by measuring differences in model configurations, performance trade-offs, and scenario sensitivity.

Similarly, Zhang, et al. (2024b) propose an architecture that incorporates an agent-based simulation model at its core to replicate the physical space and integrates interpretable machine learning with goal modeling to explain autonomous decisions to human operators within a DT system. Unlike black-box models, interpretable machine learning models (e.g., decision trees, k-nearest neighbors, and support vector machines) are preferred for their inherent interpretability, as they provide transparency into their decision-making processes. Human operator feedback is strategically integrated at critical stages of the modeling process with the use of controllers at the following stages Sensor Re-configurator, Model Updater, and Behavior Optimizer. This feedback loop interaction ensures that the system evolves with human oversight, maintaining alignment with operational requirements, and ethical standards. This architecture offers three types of explanations for autonomous decisions. First, measurement adaptation focuses on why and how data collection is adjusted, with benefits including improved state estimation or cost savings. Second, model adaptation explains changes to the model itself, such as parameter calibration or knowledge updates, to enhance fidelity. Finally, system behavior adaptation involves explaining what-if scenarios and how they help optimize system behavior, linking to design-phase requirements or performance metrics.

3.3.4 On Self-Explainability Frameworks

The work by Michael et al. (2024) focuses on enhancing the self-explainability of DTs in cyber-physical systems through the proposed Monitor, Analyze, Build, and Explain (MAB-EX) framework. This model-driven approach leverages various formal models — including system, process, and reasoning models—to generate explanations derived from their underlying simulation logic. The MAB-EX framework is structured to create a multi-level tree of explanations, allowing stakeholders to access varying levels of detail based on their expertise. This hierarchical design balances accessibility for non-technical users while providing in-depth insights for specialists.

It is, however, important to note that while the framework automates explanation generation, it does not include feedback loops for refining system behavior based on user input. So, whereas human-in-the-loop approaches—where stakeholders' feedback directly refines models, validates decisions, or adapts model behavior— Michael et al.'s method focuses on automated, one-way explanation generation.

Additionally, deeper explanation layers require parsing complex dependencies between system models, which can strain real-time performance in dynamic environments. During the "Monitor" phase, the DT continuously tracks system and environmental data to detect critical events, such as anomalies or significant state transitions. To optimize computational efficiency, it prioritizes high-impact events based on predefined metrics, such as safety risks and performance deviations. This targeted approach ensures real-time analysis and explanation generation while minimizing unnecessary processing, allowing the system to remain responsive and efficient, even in resource-constrained environments.

4 CHALLENGES AND OPEN QUESTIONS

As outlined in the previous sections, explainability is gaining attention in the context of DTs, yet significant challenges remain in achieving transparent, trustworthy, and actionable explanations for stakeholders. In the following, we elaborate on the key challenges and open research questions, emphasizing gaps that hinder the practical deployment of explainable simulation-based DTs.

Complexity of DT architectures: As previously discussed in Section 3.1, the emergent behavior arising from simulations can be opaque, making it difficult for users to understand the relationship between input parameters and resulting outcomes. This is further compounded by DTs' connections to real-time data streams and their capacity for autonomous decision-making. This complexity makes it impractical to trace the logic behind specific results without a thorough understanding of the underlying model design. Consequently, the trustworthiness of decisions made by DTs can be undermined by a lack of explainability. Thus, the research question that arises is how can explainability be integrated across the full lifecycle of DTs, from data collection to decision support.

Balancing computational efficiency and explainability: Achieving a high level of explainability can come at the cost of increased computational resources and time, while optimizing for efficiency might lead to less transparent models. This can force trade-offs between the level of detail and computational efficiency (Giabbanelli 2024). The Shapley values technique, while effective in quantifying feature importance, evaluate the performance of the model across all possible subsets of features, leading to high computational costs that scale exponentially with input dimensionality (Riis et al. 2022). Moreover, generating explanations at runtime can introduce additional computational complexity, which is particularly relevant for DTs running on resource-constrained edge devices (Michael et al. 2024).

Real-time constraints and explainability enables stakeholders to act on insights with minimal latency. DTs demand continuous data ingestion and iterative model updates to reflect real-world conditions, leading to delays. A key challenge lies in balancing the need for timely decisions with the criticality of explanations. In safety-critical applications, speed often takes precedence over detailed explanations, while a thorough analysis afterward is preferred for gaining comprehensive insights. Li et al. (2020) proposed an adaptive cost-aware approach that uses probabilistic model checking to determine when to provide explanations, optimizing the timing based on risk levels and operational context. This method highlights that while explanations can increase the probability of human operators successfully completing tasks, they may also introduce comprehension delays that need to be managed. Zhang et al. (2024b) explored this concept in the context of human-in-the-loop DTs, leveraging similar principles to optimize explanation timing. Furthermore, the integration of LLMs in DTs introduces additional challenges: although LLMs can support explainability, their computational intensity may cause delays in response generation, ultimately impacting real-time performance (Yang et al. 2025).

Data heterogeneity, multi-modality, and integrating expert knowledge: The integration of diverse data types and modalities, such as structured data, unstructured text, images, and sensor outputs, presents a significant challenge for explainability in DTs. The varying data types come with distinct characteristics and structures, making it difficult to combine them into a unified model. For example, while structured data can be easily integrated into mathematical models, unstructured data (text, images) need specialized processing, and sensor outputs must be continuously synchronized. Additionally, incorporating expert knowledge into DT models can be complex, as it requires translating human expertise into a format that can be understood and utilized by the system. Expert knowledge needs to continuously be updated to ensure consistency (Jungmann and Lazarova-Molnar 2024). The challenge lies in ensuring that all these data types align correctly within the DT's framework and contribute to the generation of accurate, coherent explanations. Therefore, a key area to explore is the role of ontologies in the standardization of heterogeneous data semantics. By providing a unified vocabulary, ontologies can make the underlying data models more explicit and interpretable, thereby enhancing traceability and reducing ambiguity in DT's operation (Karabulut et al. 2024).

Evaluating explainability in DTs: The absence of standardized benchmarks for explainability in simulation-based models creates additional challenges. Unlike, for example, traditional models, where accuracy metrics provide clear benchmarks, the quality of explanations is inherently more subjective and contextual. Various stakeholders may have different criteria for what constitutes a good explanation (e.g., engineers might prioritize technical accuracy, managers might value actionability, and regulators might focus on compliance aspects). Currently, methods and metrics for assessing explainability in DTs are lacking. Wang et al. (2021) presented metrics such as Vmodel, Vtradeoff, and Vscenario, aiming to quantify explanation value but they fall short of providing a comprehensive comparative evaluation. The integration of LLMs for generating explanations introduces new challenges, including the risk of hallucinations, thereby necessitating thorough human expert validation (Zhang et al. 2024b; Giabbanelli 2024). Finally, insights from established methods for evaluating model explanations from explainable AI could provide valuable guidance. This gap highlights the need for evaluation frameworks that can effectively address both qualitative and quantitative aspects of explainability in DTs.

Ethical and regulatory implications: The implementation of explainable DTs, particularly those involving sensitive data or high-stakes decisions, raises crucial ethical and regulatory concerns. It is vital to ensure that explanations generated by DTs are fair, unbiased, and respect privacy. One dilemma would be that the notion of fairness and equity varies across applications of ABM frameworks (Giabbanelli 2024). Fairness often emphasizes ensuring all agents receive some benefit, even if unequally, while maintaining provider viability (Thorve et al. 2024). This contrasts with traditional approaches that average outcomes across agents, advocating for a nuanced equity assessment in ABM to address uneven benefit distributions (Steger et al. 2022). Another major ethical concern for LLMs is their potential to generate biased or harmful content (Yang et al. 2025), which necessitates ongoing inspection of their outputs. From a legal perspective, LLMs often require access to vast amounts of personal data for training, raising issues of data privacy and intellectual property rights. Therefore, to ensure that explanations mitigate bias and comply with regulations, auditability and accountability mechanisms are needed (Goodman and Flaxman 2017). As a result, the key question to explore is how explanations can balance transparency with data privacy.

Fidelity, comprehensibility, and trustworthiness: Balancing explanations fidelity with stakeholders' comprehensibility remains a challenge. Explanations must be both technically accurate and understandable to non-experts (Michael et al. 2024). This requires tailoring them to the audience's expertise, ensuring they are actionable, and fostering trust in the DT system. However, trust in DT explanations depends on both their technical quality and their alignment with users' expectations. This highlights the need for explanation approaches that bridge the gap between accuracy and human comprehension, potentially through adaptive interfaces that adjust to stakeholders' knowledge levels and information needs.

User studies and human-centered evaluation methods are essential for validating explainable DTs, yet systematic approaches remain limited. Wang et al. (2021) highlight stakeholder participation and propose a framework validated through human subject experiments, demonstrating increased user confidence and trust. Zhang et al. (2024a) further stress that explanations enable users to understand decision rationales and intervene when necessary, highlighting the critical role of user evaluations for effective interaction. Ultimately, explainability should enhance understanding and decision-making. User studies can help assess whether explanations are clear, meaningful, and relevant to the target audience, ensuring they are tailored to different user groups and lead to better outcomes.

5 SUMMARY AND OUTLOOK

This paper provides an exploration of explainability in simulation-based Digital Twins, synthesizing current advancements, persistent challenges, and critical gaps. As Digital Twins continue to evolve, their capacity to support decision-making in complex, dynamic environments will depend not only on their technical sophistication but also on their ability to communicate their reasoning in an accessible and actionable way to diverse stakeholders. Our review highlights diverse approaches to enhancing explainability, including Large Language Model-driven explanations, post-hoc interpretability methods, collaborative Human-in-the-Loop frameworks, and self-explainability frameworks. These methods are successful in addressing domain-specific needs; however, existing approaches remain fragmented, focusing on isolated aspects of Digital Twins rather than the full lifecycle, from data validation to decision support.

Two particularly promising directions for explainable Digital Twins that are of interest to us are: first, developing unified frameworks that embed explainability across the lifecycle to ensure transparency and trust throughout; and second, tackling data heterogeneity and integrating expert knowledge via ontologies to provide a consistent, interpretable structure for diverse data sources and domain insights. Addressing these gaps will require adaptive explanation systems capable of dynamically adjusting detail and modality to the user's expertise and context. Interdisciplinary collaboration between simulation experts, human-computer interaction researchers, and ethicists will be needed to balance technical rigor with usability and regulatory compliance.

ACKNOWLEDGMENTS

The authors extend their thanks for the funding received from the ONE4ALL project (Grant Agreement No. 101091877) and the DMaaST project (Grant Agreement No. 101138648), both funded by the European Commission under the Horizon Europe Programme.

REFERENCES

Abdeen, Fathima Nishara, and Samad M. E. Sepasgozar. 2021. "City Digital Twin Concepts: A Vision for Community Participation." *Environmental Sciences Proceedings* 12 (1): 19. https://doi.org/10.3390/environsciproc2021012019.

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access* 6:52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.

Ali, Zeeshan, Raheleh Biglari, Joachim Denil, Joost Mertens, Milad Poursoltan, and Mamadou Kaba Traoré. 2024. "From Modeling and Simulation to Digital Twin: Evolution or Revolution?" *SIMULATION* 100 (7): 751–69. https://doi.org/10.1177/00375497241234680.

Axelrod, Robert. 1997. "Advancing the Art of Simulation in the Social Sciences." In *Simulating Social Phenomena*, edited by Rosaria Conte, Rainer Hegselmann, and Pietro Terna, 21–40. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-662-03366-1_2.

Blasek, Nico, Karl Eichenmüller, Bastian Ernst, Niklas Götz, Benjamin Nast, and Kurt Sandkuhl. 2023. "Large Language Models in Requirements Engineering for Digital Twins." *Companion Proceedings of PoEM & EDEWC – Companion 2023.*.

Boschert, Stefan, and Roland Rosen. 2016. "Digital Twin—The Simulation Aspect." In *Mechatronic Futures: Challenges and Solutions for Mechatronic Systems and Their Designers*, edited by Peter Hehenberger and David Bradley, 59–74. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32156-1_5.

Correia, Jaqueline B., Mara Abel, and Karin Becker. 2023. "Data Management in Digital Twins: A Systematic Literature Review." *Knowledge and Information Systems* 65 (8): 3165–96. https://doi.org/10.1007/s10115-023-01870-1.

Friederich, Jonas, Deena P. Francis, Sanja Lazarova-Molnar, and Nader Mohamed. 2022. "A Framework for Data-Driven Digital Twins of Smart Manufacturing Systems." *Computers in Industry* 136 (April):103586. https://doi.org/10.1016/j.compind.2021.103586.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. "Retrieval-Augmented Generation for Large Language Models: A Survey." arXiv. https://doi.org/10.48550/arXiv.2312.10997.

Ghenai, Chaouki, Lama Alhaj Husein, Marwa Al Nahlawi, Abdul Kadir Hamid, and Maamar Bettayeb. 2022. "Recent Trends of Digital Twin Technologies in the Energy Sector: A Comprehensive Review." *Sustainable Energy Technologies and Assessments* 54 (December):102837. https://doi.org/10.1016/j.seta.2022.102837.

Giabbanelli, Philippe J. 2024. "Emerging Directions in Leveraging Machine Intelligence for Explainable and Equity-Focused Simulation Models of Mental Health." *Proceedings of the AAAI Symposium Series* 4 (1): 298–302. https://doi.org/10.1609/aaaiss.v4i1.31805.

Glomsrud, Jon, André Ødegårdstuen, Asuncion Clair, and Oyvind Smogeli. 2019. "Trustworthy versus Explainable AI in Autonomous Vessels." In . Helsinki, Finland.

Goodman, Bryce, and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation.'" *AI Magazine* 38 (3): 50–57. https://doi.org/10.1609/aimag.v38i3.2741.

Grigoryan, Gayane. 2024. "Explainable Artificial Intelligence for Simulation Models." In *Proceedings of the 38th ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, 59–60. SIGSIM-PADS '24. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3615979.3662148.

Hamm, Pascal, Michael Klesel, Patricia Coberger, and H. Felix Wittmann. 2023. "Explanation Matters: An Experimental Study on Explainable AI." *Electronic Markets* 33 (1): 17. https://doi.org/10.1007/s12525-023-00640-9.

Huang, Lei, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, et al. 2025. "A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions." *ACM Transactions on Information Systems* 43 (2): 1–55. https://doi.org/10.1145/3703155.

Jungmann, Michelle, and Sanja Lazarova-Molnar. 2024. "Towards Fusing Data and Expert Knowledge for Better-Informed Digital Twins: An Initial Framework." *Procedia Computer Science* 238:639–46. https://doi.org/10.1016/j.procs.2024.06.072.

Kamel Boulos, Maged N., and Peng Zhang. 2021. "Digital Twins: From Personalised Medicine to Precision Public Health." *Journal of Personalized Medicine* 11 (8): 745. https://doi.org/10.3390/jpm11080745.

Karabulut, Erkan, Salvatore F. Pileggi, Paul Groth, and Victoria Degeler. 2024. "Ontologies in Digital Twins: A Systematic Literature Review." *Future Generation Computer Systems* 153 (April):442–56. https://doi.org/10.1016/j.future.2023.12.013.

Koopman, Philip, and Michael Wagner. 2017. "Autonomous Vehicle Safety: An Interdisciplinary Challenge." *IEEE Intelligent Transportation Systems Magazine* 9 (1): 90–96. https://doi.org/10.1109/MITS.2016.2583491.

Law, Averill M. 2015. *Simulation Modeling and Analysis*. Fifth Edition. New York: McGraw-Hill US Higher Ed USE Legacy.

Lazarova-Molnar, Sanja. 2024. "A Vision for Advancing Digital Twins Intelligence: Key Insights and Lessons from Decades of Research and Experience with Simulation:" In *Proceedings of the 14th International Conference on Simulation and Modeling Methodologies, Technologies and Applications*, 5–10. Dijon, France: SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0012884800003758.

Li, Nianyu, Javier Cámara, David Garlan, and Bradley Schmerl. 2020. "Reasoning about When to Provide Explanation for Human-Involved Self-Adaptive Systems." In *2020 IEEE International Conference on Autonomic Computing and Self-Organizing Systems (ACSOS)*, 195–204. https://doi.org/10.1109/ACSOS49614.2020.00042.

Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *ACM Comput. Surv.* 55 (9): 195:1-195:35. https://doi.org/10.1145/3560815.

Lorscheid, Iris, Bernd-Oliver Heine, and Matthias Meyer. 2012. "Opening the 'Black Box' of Simulations: Increased Transparency and Effective Communication through the Systematic Design of Experiments." *Computational and Mathematical Organization Theory* 18 (1): 22–62. https://doi.org/10.1007/s10588-011-9097-3.

Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." arXiv. https://doi.org/10.48550/arXiv.1705.07874.

Magosi, Zoltan Ferenc, Christoph Wellershaus, Viktor Roland Tihanyi, Patrick Luley, and Arno Eichberger. 2022. "Evaluation Methodology for Physical Radar Perception Sensor Models Based on On-Road Measurements for the Testing and Validation of Automated Driving." *Energies* 15 (7): 2545. https://doi.org/10.3390/en15072545.

Maidstone, Robert. 2012. "Discrete Event Simulation, System Dynamics and Agent Based Simulation: Discussion and Comparison," March, 1–6.

Michael, Judith, Maike Schwammberger, and Andreas Wortmann. 2024. “Explaining Cyberphysical System Behavior With Digital Twins.” *IEEE Software* 41 (1): 55–63. https://doi.org/10.1109/MS.2023.3319580.

Miller, Tim. 2019. “Explanation in Artificial Intelligence: Insights from the Social Sciences.” *Artificial Intelligence* 267 (February):1–38. https://doi.org/10.1016/j.artint.2018.07.007.

Nigel, Gilbert, and Troitzsch Klaus. 2005. *Simulation For The Social Scientist*. McGraw-Hill Education (UK).

Proper, Henderik A, Dominik Bork, and Geert Poels. 2021. “Towards an Ontology-Driven Approach for Digital Twin Enabled Governed IT Management.”

Retzlaff, Carl O., Alessa Angerschmid, Anna Saranti, David Schneeberger, Richard Röttger, Heimo Müller, and Andreas Holzinger. 2024. “Post-Hoc vs Ante-Hoc Explanations: xAI Design Guidelines for Data Scientists.” *Cognitive Systems Research* 86 (August):101243. https://doi.org/10.1016/j.cogsys.2024.101243.

Riis, Christoffer, Francisco Antunes, Tatjana Bolic, Gerald Gurtner, Francisco Camara Pereira, and Carlos Lima Azevedo. 2022. “Explainable Metamodels for ATM Performance Assessment.” In *Proceedings of the 12th SESAR Innovation Days*. Hungary, Budapest.

Rodis, Nikolaos, Christos Sardianos, Panagiotis Radoglou-Grammatikis, Panagiotis Sarigiannidis, Iraklis Varlamis, and Georgios Th Papadopoulos. 2024. “Multimodal Explainable Artificial Intelligence: A Comprehensive Review of Methodological Advances and Future Research Directions.” arXiv. https://doi.org/10.48550/arXiv.2306.05731.

Shapley, L. S. 1953. “17. A Value for n-Person Games.” In *Contributions to the Theory of Games (AM-28), Volume II*, edited by Harold William Kuhn and Albert William Tucker, 307–18. Princeton University Press. https://doi.org/10.1515/9781400881970-018.

Steger, Cara, Tim Williams, Daniel Brown, Seth Guikema, Birgit Müller, Tom Logan, and Nicholas Magliocca. 2022. “Integrating Equity Considerations into Agent-Based Modeling: A Conceptual Framework and Practical Guidance.” *Journal of Artificial Societies and Social Simulation* 25 (May). https://doi.org/10.18564/jasss.4816.

Sun, Zhiyan, Sanduni Jayasinghe, Amir Sidiq, Farham Shahrivar, Mojtaba Mahmoodian, and Sujeeva Setunge. 2025. “Approach Towards the Development of Digital Twin for Structural Health Monitoring of Civil Infrastructure: A Comprehensive Review.” *Sensors* 25 (1): 59. https://doi.org/10.3390/s25010059.

Thorve, Swapna, Henning Mortveit, Anil Vullikanti, Madhav Marathe, and Samarth Swarup. 2024. “Assessing Fairness of Residential Dynamic Pricing for Electricity Using Active Learning with Agent-Based Simulation.” In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 1827–36. AAMAS ’24. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Vilone, Giulia, and Luca Longo. 2021. “Notions of Explainability and Evaluation Approaches for Explainable Artificial Intelligence.” *Information Fusion* 76 (December):89–106. https://doi.org/10.1016/j.inffus.2021.05.009.

Wang, Lu, Tianhu Deng, Zeyu Zheng, and Zuo-Jun Max Shen. 2021. “Explainable Modeling in Digital Twin.” In *2021 Winter Simulation Conference (WSC)*, 1–12. https://doi.org/10.1109/WSC52266.2021.9715321.

Wu, Xiaoliang, Peter Bell, and Ajitha Rajan. 2023. “Explanations for Automatic Speech Recognition.” 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). arXiv. https://doi.org/10.48550/arXiv.2302.14062.

Yang, Jinho, Yoo Ho Son, Donggun Lee, and Sang Do Noh. 2022. “Digital Twin-Based Integrated Assessment of Flexible and Reconfigurable Automotive Part Production Lines.” *Machines* 10 (2): 75. https://doi.org/10.3390/machines10020075.

Yang, Linyao, Shi Luo, Xi Cheng, and Lei Yu. 2025. “Leveraging Large Language Models for Enhanced Digital Twin Modeling: Trends, Methods, and Challenges.” arXiv. https://doi.org/10.48550/arXiv.2503.02167.

Zhang, Nan, Rami Bahsoon, Nikos Tziritas, and Georgios Theodoropoulos. 2024. “Explainable Human-in-the-Loop Dynamic Data-Driven Digital Twins.” In *Dynamic Data Driven Applications Systems*, edited by Erik Blasch, Frederica Darema, and Alex Aved, 233–43. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-52670-1_23.

Zhang, Nan, Christian Vergara-Marcillo, Georgios Diamantopoulos, Jingran Shen, Nikos Tziritas, Rami Bahsoon, and Georgios Theodoropoulos. 2024. “Large Language Models for Explainable Decisions in Dynamic Digital Twins.” *Proceedings of the 5th International Conference on Dynamic Data Driven Applications Systems (DDDAS 2024)*, 9. https://doi.org/10.48550/ARXIV.2405.14411.

AUTHOR BIOGRAPHIES

MERYEM MAHMOUD is a Ph.D. candidate at the Institute of Applied Informatics and Formal Description Methods at Karlsruhe Institute of Technology in Germany. Her research focuses on explainability and ontologies in Digital Twins. Her email address is meryem.mahmoud@kit.edu.

SANJA LAZAROVA-MOLNAR is a Professor at both the Karlsruhe Institute of Technology and the University of Southern Denmark. Her research focuses on data-driven simulation, Digital Twins, and cyber-physical systems modeling, with an emphasis on reliability and energy efficiency. She develops advanced methodologies to optimize complex systems and leads several European and national projects in these areas. Prof. Lazarova-Molnar holds leadership roles in IEEE and The Society for Modeling & Simulation International (SCS), where she currently serves as SCS Representative to the Winter Simulation Conference (WSC) Board of Directors. She was Proceedings Editor for WSC in 2019 and 2020 and serves as Associate Editor for *SIMULATION: Transactions of The Society for Modeling and Simulation International*. Her email address is lazarova-molnar@kit.edu.