frontiers | Frontiers in Mechanical Engineering

TYPE Perspective
PUBLISHED 05 January 2024
DOI 10.3389/fmech.2023.1329146

Check for updates

OPEN ACCESS

EDITED BY
Mrityunjay Doddamani,
Indian Institute of Technology Mandi,
India

REVIEWED BY
Shohin Aheleroff,
SUEZ Smart Solutions, New Zealand

*CORRESPONDENCE
Joseph Cohen,
✉ cohenyo@umich.edu

RECEIVED 28 October 2023
ACCEPTED 18 December 2023
PUBLISHED 05 January 2024

CITATION
Cohen J and Huan X (2024), Uncertainty-aware explainable AI as a foundational paradigm for digital twins.
Front. Mech. Eng 9:1329146.
doi: 10.3389/fmech.2023.1329146

Uncertainty-aware explainable AI as a foundational paradigm for digital twins

Joseph Cohen[1]* and Xun Huan[2]

[1]Michigan Institute for Data Science, University of Michigan, Ann Arbor, MI, United States, [2]Department of Mechanical Engineering, University of Michigan, Ann Arbor, MI, United States

In the era of advanced manufacturing, digital twins have emerged as a foundational technology, offering the promise of improved efficiency, precision, and predictive capabilities. However, the increasing presence of AI tools for digital twin models and their integration into industrial processes has brought forth a pressing need for trustworthy and reliable systems. Uncertainty-Aware eXplainable Artificial Intelligence (UAXAI) is proposed as a critical paradigm to address these challenges, as it allows for the quantification and communication of uncertainties associated with predictive models and their corresponding explanations. As a platform and guiding philosophy to promote human-centered trust, UAXAI is based on five fundamental pillars: accessibility, reliability, explainability, robustness, and computational efficiency. The development of UAXAI caters to a diverse set of stakeholders, including end users, developers, regulatory bodies, the scientific community, and industrial players, each with their unique perspectives on trust and transparency in digital twins.

KEYWORDS

explainable artificial intelligence, uncertainty quantification, machine learning, digital twins, advanced manufacturing

1 Introduction

Enabled by Industry 4.0 technologies, Digital Twins (DTs) have attracted enormous attention for advanced manufacturing in recent years. Due to recent advancements in cloud storage and computing, Artificial Intelligence (AI), and Industrial Internet of Things (IIoT)–enabled connectivity of Cyber–Physical Systems (CPS), DT systems have become increasingly realizable (Bergs et al., 2021). First introduced two decades ago and formalized by Grieves (2014), the DT concept—of which several definitions and variations have been developed since—has important implications on real-time condition monitoring of physical assets in manufacturing production systems. For clarity, this work will use the DT definition used by Kochunas and Huan (2021), which considers the existence of the DT as tied to the active operation of a physical asset in the context of its life cycle. This contrasts with other digital models and digital shadows (Bergs et al., 2021) that may exist before or after the monitored asset is in-service. As elaborated by Kochunas and Huan (2021), digital models, twins, and shadows are all valid aspects that exist within the entire life cycle of an asset, including after it is decommissioned.

Overall, while the DT concept is relatively new in implementation, DTs are largely based upon existing and mature modeling and simulation components (Fuller et al., 2020). We refer to Aheleroff et al. (2021) for more information concerning enabling technologies for DTs and the value that existing DTs provide for specific industries. Modeling components of a DT system may consist of a combination of experience-based, physics-based, data-driven models, and hybrid approaches combining these modes (Liao and Köttig, 2016). Experience-based techniques are often heuristic rules carefully designed by the accumulation of extensive human expert knowledge (Liao and Köttig, 2016). Physics-based simulations may be too computationally expensive to resolve in real time, whereas data-driven models can perform tasks such as prediction, inference, and control in online settings once trained. Due to the recent development and feasibility of Machine Learning (ML) techniques, data-driven models have become increasingly popular alternatives for real-time monitoring that is integral for DTs (Jaensch et al., 2018), particularly in advanced manufacturing applications where novel and complex physics are difficult to fully capture. DT systems may be employing ML for an assortment of classification or regression tasks, and reinforcement learning techniques to obtain optimal policies for decision-making. For example, Alexopoulos et al. (2020) proposed a DT framework using deep learning for vision-based inspection. Their proposal included simulating virtual datasets to streamline manual labeling efforts. Meanwhile, Xia et al. (2021) demonstrated the utility of deep reinforcement learning for operations optimization tasks, including manufacturing scheduling.

Due to these recent developments, DTs increasingly depend upon black box predictive models. However, the opaqueness of existing ML methods creates obstacles for trust, inhibiting potential for DT adoption in industry. This can have devastating consequences for high-stakes and/or safety-critical applications where trustworthiness is a priority. Defining trustworthiness in the context of DTs is itself nontrivial. Doroftei et al. (2021) proposed 4 dimensions of human-agent trust for DTs: reliable performance, process understanding, intended use conforming to design purpose, and socio-ethical implications. Similarly, Trauer et al. (2022) proposed a 7-step Trust Framework for DTs after surveying industry professionals. In a recent report, the National Institute of Standards and Technology (NIST) discussed the emerging DTs in the context of existing standards and guidelines, offering considerations on trustworthy applications (Voas et al., 2021). Despite these contributions, defining and establishing trustworthiness for DTs in a systematic, comprehensive, and context-relevant manner remains a major challenge.

This perspective paper will first discuss gaps in DT trustworthiness through a variety of stakeholder perspectives, elucidating current factors limiting confidence in the data-driven models that are the building blocks of DT operation. The paper will then delineate five core pillars of trustworthy DTs to address these limitations, and how the development of uncertainty-aware explainable artificial intelligence (UAXAI) as an underlying fundamental framework can facilitate the adoption of explainable and trustworthy DT systems. The main contribution of this paper is then to, through the presentation of UAXAI, provide structure and guidance for upcoming future research efforts to better streamline the transition towards resilient, sustainable, and human-centered AI in Industry 5.0.

2 Stakeholder perspectives on trustworthiness

To facilitate the adoption of trustworthy data-driven DTs, this paper considers five diverse stakeholder perspectives: end user, developer, regulatory, scientific, and industrial. The proposed UAXAI framework, introduced in the next section, is motivated from all the perspectives elaborated in this section.

2.1 End user perspective

From an end user standpoint, trust in DTs is paramount. Operators and decision-makers rely on these systems for real-time insights, and the reliability of predictions can directly affect manufacturing production. Black box models often give point estimates with no explanation or justification for predictions, with no uncertainty or confidence measure. This can inhibit trust for end users, who may appreciate more information to make judgments on whether to trust the prediction and health status of the DT operation. An important challenge is that state-of-the-art accurate and eXplainable Artificial Intelligence (XAI) methods characterized as “trustworthy” may still easily deceive end users. In a recent study in human-computer interaction, Banovic et al. (2023) demonstrated how exaggerating the capabilities of an untrustworthy AI system can obscure end users’ objectivity in evaluating its reliability. This demonstrates that our criteria for evaluating user trust cannot be limited to simply gauging the apparent quality and fairness of predictive models used for DTs.

2.2 Developer perspective

Developers face the challenge of building and maintaining DT systems that are robust and efficient. With current black box techniques, it is difficult for developers to identify regions of strength and weakness in the model’s predictions. Data-driven methods are also often incompatible with unlabeled data ubiquitous in manufacturing applications. In addition, they can be exceedingly difficult to train and tune due to challenges in convergence: for example, the loss function when training a neural network tends to be highly nonconvex, and trainable parameters are often stuck in local minima with no guarantee on global optimality (von Eschenbach, 2021). Additionally, performance often relies on data-dependent and opaque hyperparameter settings.

2.3 Regulatory perspective

Regulatory bodies are increasingly concerned with the ethical and safe deployment of AI systems. The recent NIST report on emerging standards for DTs brought forth 14 considerations defining trust as “the probability that the intended behavior and the actual behavior are equivalent given a fixed context, fixed environment, and fixed point in time” (Voas et al., 2021). Compliance with legacy standards and guidelines is especially important for safety-critical systems, where failure could have catastrophic implications on lives and costs. Purely data-driven models may be exceedingly brittle for worry-free usage. This also manifests in DTs: small perturbations in operating conditions may lead to divergent and unreliable predictions. Without developed safeguards and engineering redundancies in place, it is difficult to guarantee system performance given unpredictable model tendencies. The development of new regulations and policies for AI systems is of intense interest, and it remains to be seen how this will impact DTs in the coming generation.

2.4 Scientific perspective

Domain scientists and engineers who have accumulated years of expertise may be skeptical of data-driven predictions from DT systems. Data-driven findings, particularly without justification or explanation, may appear to contradict established science and convention. In general, ascertaining whether these findings are indicative of physical phenomena or are simply experimental artifacts remains nontrivial. While skepticism under uncertainty is important for robust decision-making, black box techniques exacerbate existing distrust and aversion to data-centric methods (von Eschenbach, 2021).

2.5 Industrial perspective

Ultimately, industry practitioners and executives are seeking solutions that improve key performance indicators such as yield, throughput, reduced downtime, availability, quality, and cost reduction. Data-driven methods for DTs can come with steep implementation costs (e.g., sensor installation, data collection and storage, GPUs/TPUs), and also come with pressing questions on cybersecurity and privacy preservation. While there have been successful case studies of DTs operating at scale (Betti et al., 2020), the decision to trust DT models from the perspective of industry largely depends on how they materially benefit business operations. Like other technologies, the adoption of DTs may resemble Gartner's Hype Cycle characterized as follows: the initial technology trigger, the peak of inflated expectations, the trough of disillusionment, the slope of enlightenment, and the plateau of productivity (Strawn, 2021). We note that this cycle is especially pertinent due to the recent explosion of generative AI capabilities such as large language models; some industry practitioners may have raised expectations due to hype, and it may take more time to understand how to deploy DT and AI technologies productively.

3 UAXAI: Uncertainty-Aware eXplainable AI

Based on the existing challenges identified above, we suggest five guiding principles for the development of trustworthy DTs: accessibility, reliability, explainability, robustness, and computational efficiency. These principles are defined in this section and will form the foundational pillars of UAXAI, a proposed framework centered on improving trustworthiness for data-driven DTs. The UAXAI framework is not limited to a specific methodology, but is characterized as a platform that incorporates human explanations and expertise. Figure 1 summarizes the framework's prioritization of each foundational concept as well as the resulting benefits towards trustworthiness.

3.1 Accessibility and UAXAI

In this context, accessibility is defined as the degree to which human operators can understand how to use and interface with DTs. This can encompass a wide variety of aspects ranging from having understandable parameter settings that are intuitive for experts to tune, to inclusive universal design principles. Data visualization tools and adaptive interfaces are essential to promote accessibility (Todi et al., 2020). Our position is that Uncertainty Quantification (UQ) and communication is a key enabler for end users to understand the risks involved with trusting a DT system with confidence. Prabhudesai et al. (2023) provided empirical evidence that including uncertainty information into decision support systems combats overreliance and promotes critical understanding. In general, the principle of accessibility can be quantified via operator survey feedback and compliance to standards, and DT interfaces should be designed to meet target accessibility specifications.

Accessibility in DTs is closely tied to human psychology, cognitive processes, and even local culture. The UAXAI framework respects the research contributions from the human-computer interaction (HCI) community, recognizing the importance of presenting information in a format that resonates with users. Unlike black box models, the UAXAI concept aims to create user-friendly AI systems that empowers operators, lowering the barrier of entry of using these techniques. By making it so that operators can easily interact and interface with DTs, practitioners can introduce these systems in reskilling programs with less friction.

3.2 Reliability and UAXAI

This principle is defined as the ability to demonstrate high performance on a quantifiable, consistent, and reproducible basis. Reliability metrics for fault diagnosis problems may include classification performance evaluation quantities such as precision, recall, area under the receiver operating characteristics (AUROC) or precision-recall (AUPR) curves. These are often more robust metrics than accuracy alone, which can easily appear misleadingly high under class imbalance situations commonly encountered in manufacturing applications.

Reliability is a fundamental concern for DTs. Users must have confidence in the predictions generated by these systems. The proposed UAXAI framework contributes to reliability by enabling the explanation of predictions, allowing accuracy metrics to be reported alongside uncertainty measures. This combination provides users with a comprehensive view of model performance. In applications with weakly labeled data, where ground truth validation is challenging, UAXAI leverages uncertainty and explanations as surrogate error and evaluation measures. It helps identify the strengths and weaknesses of the model, allowing for continuous improvement and enhanced decision-making.

3.3 Explainability and UAXAI

This principle focuses on transparency, the ability to understand and trust predictions, as well as the long-term implications of decision-making. In this paper, explainability is considered a central pillar for establishing trust. While explainability as a concept is difficult to objectively measure, feature importance and attribution methods such as Shapley value analysis are quantitative tools that can be used to explain model predictions (Senoner et al., 2021). Example-based explanations and model counterfactuals (i.e., "the smallest possible changes to produce a different outcome") are particularly amenable for human learning and intuitive from a HCI perspective.

However, existing XAI techniques have significant limitations that can lead to misleading explanations. Feature attribution techniques such

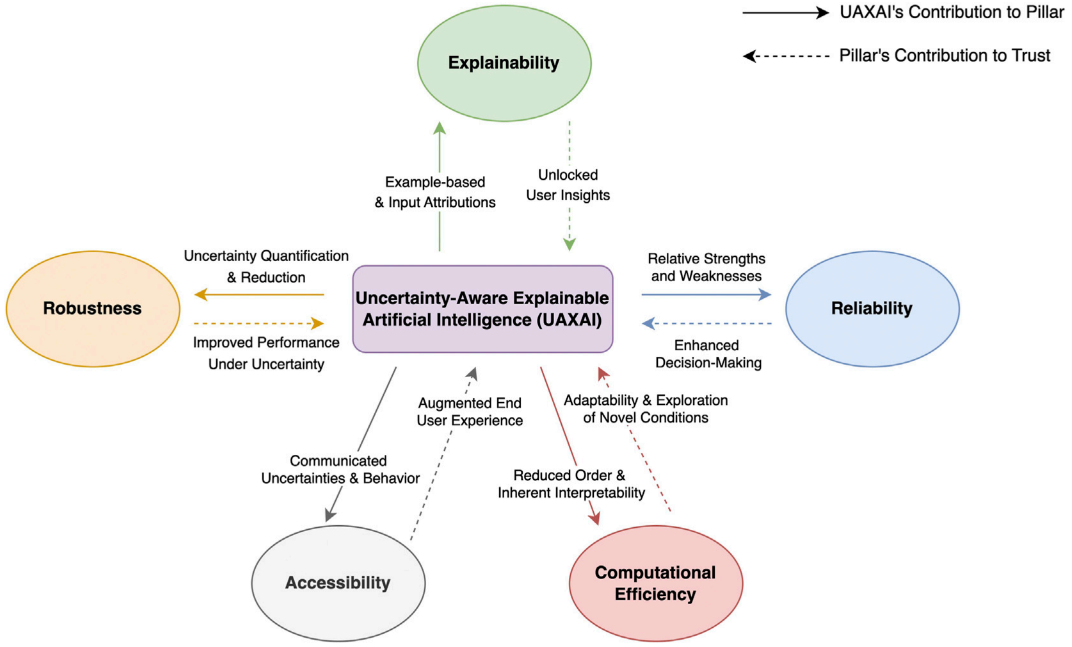
UAXAI's Contribution to Pillar
Pillar's Contribution to Trust
Explainability
Example-based & Input Attributions
Unlocked User Insights
Uncertainty Quantification & Reduction
Robustness
Improved Performance Under Uncertainty
Uncertainty-Aware Explainable Artificial Intelligence (UAXAI)
Relative Strengths and Weaknesses
Reliability
Enhanced Decision-Making
Augmented End User Experience
Adaptability & Exploration of Novel Conditions
Communicated Uncertainties & Behavior
Reduced Order & Inherent Interpretability
Accessibility
Computational Efficiency

FIGURE 1
The proposed conceptual UAXAI framework based on five pillars: accessibility, reliability, explainability, robustness, and computational efficiency.

TABLE 1 Summary table matching identified stakeholders' unmet needs with value provided by UAXAI platform. We refer back to Figure 1 for a summary on how these pillars contribute to trustworthiness.

Perspective	Summary of unmet challenges	Relevant UAXAI pillars
End User	Inability to understand or justify predictions	Explainability, Accessibility
Developer	Difficulty in identifying model strengths and weaknesses	Explainability, Reliability
Regulatory	Assuring compliance to existing standards, societal responsibility	Robustness, Reliability
Scientific	Skepticism in opaque and black–box model results	Explainability, Computational Efficiency
Industrial	Skepticism in improving key performance indicators	Reliability, Computational Efficiency

as Shapley-based methods have a variety of estimation techniques and varying interpretations of marginal effects (Chen et al., 2023). Other data-dependent techniques, such as Local Interpretable Model-Agnostic Explanations (LIME), are sensitive to implementation details and prone to unstable, inconsistent, and non-unique explanations (Molnar, 2022). As a result, without proper communication of uncertainties and context, XAI methods can be deceptive. A UAXAI platform should incorporate these tools appropriately by unlocking user insights on DT behavior, allowing for a more honest assessment on relative strengths and weaknesses.

3.4 Robustness and UAXAI

This principle assures capability in handling uncertainty, noise, and disturbances to provide meaningful, actionable solutions. For data-driven methods, forward UQ can be useful as a robustness measure to evaluate uncertainty in predictions (Kochunas and Huan, 2021). Furthermore, inverse UQ can be utilized alongside new data to reduce uncertainty. This can be accomplished via Bayesian or frequentist approaches. Bayesian frameworks allow for a more comprehensive overview in terms of probabilistic distributions, whereas frequenstist approaches can be useful to swiftly calculate confidence measures.

DTs are expected to operate under various conditions, including changing environments and noisy sensor data. UAXAI can play a crucial role in ensuring robustness by quantifying and propagating uncertainty. It can account for noisy sensor measurements in real-time, making DTs resilient to fluctuations. Moreover, UAXAI serves as an enabler for resilience, alerting operators when model explanations no longer align with observed data. This indication can help encourage retraining or remodeling, ensuring that the DT remains effective and trustworthy.

3.5 Computational efficiency and UAXAI

This principle places careful consideration on runtime and computational cost, including offline and online preprocessing, postprocessing, training, evaluation, and retraining steps. Some examples of quantifiable computational efficiency metrics include total runtime, number of iterations required for convergence, and time complexity. DT specifications can be targeted on a per-application basis to reach these targets, with stringent requirements necessary to achieve real-time decision-making.

Efficiency is a key concern in the deployment of DTs. UAXAI addresses this by advocating for multi-fidelity approaches, where expensive high-fidelity physics-based simulations are complemented with low-fidelity reduced-order models (Kapteyn et al., 2020), or inherently interpretable models that are computationally efficient. It also supports optimal experimental design, aiding in the cost-effective collection of data. Rather than relying on a single large, high-fidelity model, UAXAI promotes a network of different models with varying fidelity levels used for different tasks. This approach allows experts to select the most appropriate model for a given situation, enhancing efficiency and effectiveness for faster exploration of vast solution and design spaces.

4 Discussion

The UAXAI framework exhibits an underlying philosophy of prioritizing the quantification and communication of uncertainties alongside model explanations. It directly addresses several of the unmet needs from the stakeholder perspectives identified in Section 2, with the overall relationship summarized in Table 1. In Table 1, we specify the two most relevant and value–adding UAXAI pillars that correspond to the respective stakeholder's needs.

The proposed framework is not limited to specific methodologies or model fidelities, but the combination of several existing technologies may fit well within this framework. For example, low-fidelity models, whether they be reduced-order models from high-fidelity simulations (Kochunas and Huan, 2021) or "inherently interpretable" linear and/or sparse data-driven models, are computationally efficient and may be more resilient to uncertainty due to generalization capability. However, an advanced deep learning algorithm that offers high accuracy, or a tool that provides accessibility benefits by unlocking savings in labeling costs for vision-based systems such as Segment Anything (Kirillov et al., 2023) also have their place in this vision. Model-agnostic (e.g., LIME and SHAP) and model-specific feature attribution methods in addition to example-based (e.g., counterfactual reasoning) explanation methods may each prove useful, especially when provided with sufficient context to avoid misleading the end user. For example, explanation uncertainty must be explored and quantified to critically evaluate the trustworthiness of model explanation methods (Cohen et al., 2023). The unifying philosophy of UAXAI as a platform harmonizes these seemingly disparate techniques to center uncertainty quantification and communication in conjunction with explaining model behavior in an accessible manner, while respecting computational and industrial limitations.

In the context of advanced manufacturing, the development of UAXAI as a foundational paradigm for DTs holds great promise. By directly addressing and valuing the perspectives of end users, developers, regulatory bodies, the scientific community, and industrial players, the proposed framework promotes transparency and human experience as important vehicles to enhance trust. An example of an early adopter of this framework could include manufacturing for aerospace applications, where high–stakes and safety–critical engineering needs are ubiquitous. We refer to Li et al. (2022) for a review of DTs for the aerospace sector. Future work must work on seamlessly integrating UQ, XAI, and HCI components to fully realize UAXAI as a framework for DTs. Ultimately, obtaining and maintaining human trust is deeply psychological, and can never be guaranteed. To this end, industry practitioners should invest on improving the synergy between human operators and data-driven DTs, promoting human-centered augmented intelligence as opposed to positioning data-driven models as adversarial expert systems that replace human intuition.

The five pillars of accessibility, reliability, robustness, computational efficiency, and explainability collectively contribute to the trustworthy adoption of advanced computation in DTs and smart manufacturing. With maturing UAXAI techniques, end users and other stakeholders can make better decisions on whether they should trust DT models given calculated risks from real-time system operation. Trustworthy DTs equipped with UAXAI will not only enhance operational efficiency, but also serve as reliable decision support tools. Going into the new Industry 5.0 era, it is crucial for stakeholders to work collaboratively to integrate UAXAI into DT ecosystems, thereby unlocking the full potential of advanced manufacturing.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

JC: Writing–original draft, Writing–review and editing. XH: Writing–original draft, Writing–review and editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This project is supported by the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Aheleroff, S., Xu, X., Zhong, R. Y., and Lu, Y. (2021). Digital twin as a service (DTaaS) in industry 4.0: an architecture reference model. *Adv. Eng. Inf.* 47, 101225. doi:10.1016/j.aei.2020.101225

Alexopoulos, K., Nikolakis, N., and Chryssolouris, G. (2020). Digital twin-driven supervised machine learning for the development of artificial intelligence applications in manufacturing. *Int. J. Comput. Integr. Manuf.* 33, 429–439. doi:10.1080/0951192X.2020.1747642

Banovic, N., Yang, Z., Ramesh, A., and Liu, A. (2023). Being trustworthy is not enough: how untrustworthy artificial intelligence (AI) can deceive the end-users and gain their trust. *Proc. ACM Hum.-Comput. Interact.* 7, 1–17. doi:10.1145/3579460

Bergs, T., Gierlings, S., Auerbach, T., Klink, A., Schraknepper, D., and Augspurger, T. (2021). The concept of digital twin and digital shadow in manufacturing. *Procedia CIRP* 101, 81–84. doi:10.1016/j.procir.2021.02.010

Betti, F., de Boer, E., and Giraud, Y. (2020). *Industry's fast-mover advantage: enterprise value from digital factories*. Atlanta, GA, USA: World Economic Forum and McKinsey & Company.

Chen, H., Covert, I. C., Lundberg, S. M., and Lee, S.-I. (2023). Algorithms to estimate Shapley value feature attributions. *Nat. Mach. Intell.* 1, 590–601. doi:10.1038/s42256-023-00657-x

Cohen, J., Byon, E., and Huan, X. (2023). To trust or not: towards efficient uncertainty quantification for stochastic Shapley explanations. *PHM Soc. Asia-Pacific Conf.* 4. doi:10.36001/phmap.2023.v4i1.3694

Doroftei, D., De Vleeschauwer, T., Bue, S. L., Dewyn, M., Vanderstraeten, F., and De Cubber, G. (2021). "Human-agent trust evaluation in a digital twin context," in 2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN), Vancouver, BC, Canada, 08-12 August 2021, 203–207. doi:10.1109/RO-MAN50785.2021.9515445

Fuller, A., Fan, Z., Day, C., and Barlow, C. (2020). Digital twin: enabling technologies, challenges and open research. *IEEE Access* 8, 108952–108971. doi:10.1109/ACCESS.2020.2998358

Grieves, M. (2014). Digital twin: manufacturing excellence through virtual factory replication. *White Pap.* 1–7.

Jaensch, F., Csiszar, A., Scheifele, C., and Verl, A. (2018). "Digital twins of manufacturing systems as a base for machine learning," in 2018 25th International conference on mechatronics and machine vision in practice (M2VIP), Stuttgart, Germany, 20-22 November 2018 (IEEE), 1–6. doi:10.1109/M2VIP.2018.8600844

Kapteyn, M. G., Knezevic, D. J., and Willcox, K. (2020). Toward predictive digital twins via component-based reduced-order models and interpretable machine learning. In *AIAA Scitech 2020 Forum*. 0418. doi:10.2514/6.2020-0418

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. arXiv:2304.02643. doi:10.48550/arXiv.2304.02643

Kochunas, B., and Huan, X. (2021). Digital twin concepts with uncertainty for nuclear power applications. *Energies* 14, 4235. doi:10.3390/en14144235

Li, L., Aslam, S., Wileman, A., and Perinpanayagam, S. (2022). Digital twin in aerospace industry: a gentle introduction. *IEEE Access* 10, 9543–9562. doi:10.1109/ACCESS.2021.3136458

Liao, L., and Köttig, F. (2016). A hybrid framework combining data-driven and model-based methods for system remaining useful life prediction. *Appl. Soft Comput.* 44, 191–199. doi:10.1016/j.asoc.2016.03.013

Molnar, C. (2022). *Interpretable machine learning: a guide for making black box models explainable*. 2 edn.

Prabhudesai, S., Yang, L., Asthana, S., Huan, X., Liao, Q. V., and Banovic, N. (2023). Understanding uncertainty: how lay decision-makers perceive and interpret uncertainty in human-AI decision making. *Proc. 28th Int. Conf. Intelligent User Interfaces*, 379–396. doi:10.1145/3581641.3584033

Senoner, J., Netland, T., and Feuerriegel, S. (2021). Using explainable artificial intelligence to improve process quality: evidence from semiconductor manufacturing. *Manag. Sci.* 68, 5704–5723. doi:10.1287/mnsc.2021.4190

Strawn, G. (2021). Open science and the hype cycle. *Data Intell.* 3, 88–94. doi:10.1162/dint_a_00081

Todi, K., Vanderdonckt, J., Ma, X., Nichols, J., and Banovic, N. (2020). AI4AUI: workshop on AI methods for adaptive user interfaces. *Proc. 25th Int. Conf. Intelligent User Interfaces Companion*, 17–18. doi:10.1145/3379336.3379359

Trauer, J., Schweigert-Recksiek, S., Schenk, T., Baudisch, T., Mörtl, M., and Zimmermann, M. (2022). A digital twin trust framework for industrial application. *Proc. Des. Soc.* 2, 293–302. doi:10.1017/pds.2022.31

Voas, J., Mell, P., and Piroumian, V. (2021). Considerations for digital twin technology and emerging standards. *Tech. Rep.* National Institute of Standards and Technology. doi:10.6028/NIST.IR.8356-draft

von Eschenbach, W. J. (2021). Transparency and the black box problem: why we do not trust AI. *Philosophy Technol.* 34, 1607–1622. doi:10.1007/s13347-021-00477-0

Xia, K., Sacco, C., Kirkpatrick, M., Saidy, C., Nguyen, L., Kircaliali, A., et al. (2021). A digital twin to train deep reinforcement learning agent for smart manufacturing plants: environment, interfaces and intelligence. *J. Manuf. Syst.* 58, 210–230. doi:10.1016/j.jmsy.2020.06.012