An Overview on the Explainability of Cyber-Physical Systems

Sanjiv S. Jha
University of St.Gallen
St. Gallen, Switzerland

Abstract

The increase in automating complicated physical processes using Cyber-Physical Systems (CPS) raises the complexity of CPS and their behavior. It creates the necessity to make them explainable. The popular Explainable Artificial Intelligence (XAI) methodologies employed to explain the behavior of CPS usually overlook the impact of physical and virtual context when explaining the outputs of decision-making software models, which are essential factors in explaining CPS' behavior to stakeholders. Hence in this article, we survey the most relevant XAI methods to identify their shortcomings and applicability in explaining the behavior of CPS. Our main findings are *(i)* Several papers emphasize the relevance of context in describing CPS. However, the approaches for explaining CPS fall short of being context-aware; *(ii)* the explanation delivery mechanisms use low-level visualization tools that make the explanations unintelligible. Finally *(iii)*, these unintelligible explanations lack actionability. Therefore, we propose to enrich the explanations further with contextual information using Semantic Technologies, user feedback, and enhanced explanation visualization techniques to improve their understandability. To that end, context-aware explanation and better explanation presentation based on knowledge graphs might be a promising research direction for explainable CPS.

1 Introduction

Cyber-Physical Systems (CPS) are capable of integrating physical and virtual processes (Lee 2008). The ability of CPS to interact with, and increase the capabilities of the physical entities through computation, communication, and control, are the facilitators of future technological advancements (Baheti and Gill 2011). However, the increasing complexity of software, hardware, and communication mechanisms may reduce the understandability and trust of stakeholders (users, auditors, engineers) in the CPS. Recent studies (Daglarli 2021) have argued in favor of providing explanations to increase user trust in such complex automation processes. Moreover, (Shin 2021) highlights the importance of causability [1] of the delivered explanations to increase the trust of users from a human factors perspective, and (Doshi-Velez et al. 2017) argues that explanations are required to establish the system's accountability on legal grounds. Understanding the behavior of the CPS not only would help to reason about the anomalous behavior of the CPS, but also allow users to take control over the ever-growing complexity of the CPS due to the increased interaction with their context. Hence, we argue that it will become essential for CPS to be able to explain their behavior to users while considering their context.

CPS may explain their behavior by themselves or through the help of external explanation methods. To become *self-explainable*, a system would need to know its working environment, internal states, user profiles, and the interactions between its software and physical components. Since there is no such CPS available yet that could explain themselves, a framework introduced as Monitor-Analyze-Build-Explain (MAB-EX) (Blumreiter et al. 2019) seems to be a promising framework for the development of self-explainable CPS in the future. Alternatively, *model-agnostic explanation* systems can help explain the behavior of a deployed CPS based on their historical data without interrupting the traditional engineering processes of a CPS. However, numerous recent studies have agreed on some drawbacks of using XAI techniques directly in this domain. For example, (Schlegel et al. 2019) deduced that commonly used XAI methods that are popular for behavioral analysis of CPS do not offer good understandability for time-series data through overlaying the feature importance on a time scale. Furthermore, (Weber and Wermter 2020) addresses the poor degree of understandability offered to explainees through feature visualization in XAI approaches due to the lack of information delivered. Since time may be utilized as a common factor in a CPS to synchronize the monitoring of the CPS' behavior and the behavior of other components in its context (Shrivastava et al. 2016), the explanation systems should consider the time-continuous behavior of the CPS when explaining them. Apart from a scarcity of time-series explanation visualization methods, the existing XAI methods are often unintelligible due to a lack of contextual awareness, user profile, and usage history. It is fair to assume that most of the work in XAI only adheres to the view of the developers about what constitutes a 'good' explanation (Miller 2019). Hence, user profiling and explanation customization will assist various users, such as managers, auditors, and maintainers, in understanding the

[1] Causability justifies for what and how something should be explained. It helps determine the relative importance of the properties of explainability.

system's behavioral logic. We argue that the customized explanations delivered through enhanced visualization techniques will increase users' understandability of the behavior of a CPS. Therefore, we aim to answer the following overarching research question (RQ) through this article:

RQ: What approaches of today's explanation systems may be applied to better explain the behavior of CPS, and what are the characteristic of CPS that need to be considered in explanation systems that are adapted for CPS?

To answer the RQ, we review the work done for explainable CPS and XAI in Section 2. Following that, we propose potential extensions and amendments to these explanation systems in Section 3 to enable intelligible and actionable explanations for the behavior of CPS.

2 Related Work

This review aims at exploring the literature that surrounds XAI and explainable CPS following the literature review guidelines by (Kitchenham and Charters 2007). To answer the RQ, we selected relevant search terms to form a search query as *explainable Cyber-physical Systems OR Explainability OR "Explainable Cyber-Physical Systems" AND XAI OR "Explainable Artificial Intelligence"*. This search yielded 900 results for the years 2015-2022 on Google Scholar (comparatively more comprehensive academic search engine (Gusenbauer 2019)). Based on the filters applied to include the unique, peer-reviewed work and content scanning of the articles for prototypical explanation methodologies used by the found results, 120 articles stood out, forming the base for our review. This section expands on the findings of the review process by discussing state-of-the-art XAI approaches for CPS applications, context-awareness, and visualization methods employed by the explanation methods.

XAI Methods for Explainable CPS

There are many supervised machine learning-based explanation systems available. However, CPS produce a massive amount of unlabeled data that are not useful for creating supervised learning-based explanations. In that regard, (Wickramasinghe et al. 2021) provides an explainable clustering approach based on a self-organizing map for generating global and local explanations from unlabeled data. Global explanations help understand an XAI system as a whole (Kopitar et al. 2019), while local explanation methods describe a single instance involving a smaller group of features. Local explanations are thus more limited in scope but typically lead to better understanding of the feature contributions than global explanation methods (Kopitar et al. 2019). Explanations about the behavior of a (cyber-physical) systems are either *intrinsic* when they can be produced from the inner states and algorithms of that system (Weber and Wermter 2020) through various explanation techniques, or *extrinsic* when given to the CPS from some external entities. For example, user feedback can be a form of extrinsic explanations. A further categorization of the popular XAI methods can be seen in Table 1. In terms of comprehending causality, (Gilpin et al. 2018) points out that users are only satisfied with explanations when the crucial '*Why?*' and '*Why not?*' questions are answered. In the CPS domain, the key to answering these questions are methods that permit the derivation of *causal understanding* of the relationships between the behavior of a CPS and factors that affect it. (Richens, Lee, and Johri 2020) argues that using counterfactual-based algorithms increases the accuracy in detecting causal factors in medical diagnosis offered using machine learning and medical CPS. Hence, counterfactual explanations may aid in the comprehension of explanations by demonstrating causal relationships between the CPS and the factors influencing CPS' behavior.

Table 1: Categorization of popular XAI methods

Scope	Approach	Explanation Methods
Intrinsic (Local and Global) (Weber and Wermter 2020)	Model-agnostic	LIME (Ribeiro, Singh, and Guestrin 2016) , SHAP (Lundberg et al. 2020), Deep Learning Important FeaTures (DeepLIFT) (Shrikumar, Greenside, and Kundaje 2017), Layer-wise Relevance Propagation (Bach et al. 2015), Counterfactual Explanations (Molnar 2020)
	Model-specific	Guided Back-propagation (Springenberg et al. 2014), Integrated Gradients (Sundararajan, Taly, and Yan 2017)
Extrinsic (Weber and Wermter 2020)	Feedback	Active or Passive User Input (Haque, Aziz, and Rahman 2014)

Contextual Influences of CPS

The cause of behavioral anomalies due to the effect of context (virtual as well as physical phenomena e.g., weather, air pressure, vibrations, latency) of CPS have been studied for many years. It has been quantified using *test chambers* and recently, using several automated data monitoring and analysis-based techniques like feature extraction with limit checking, clustering, and Knowledge-based methods (Lopez et al. 2017; Ricard and Owezarski 2020). To explicitly consider the context of the CPS for the reasoning of the CPS' behavior, recent research works (Sahlab, Jazdi, and Weyrich 2020) (Petnga and Austin 2013) leverage semantic technologies to model and scope the context of the CPS often using expert knowledge. Similarly, (Aryan et al. 2021) gives an example of explainable CPS using ontologies and expert knowledge. For explaining a complex scenario of demand response in a smart grid, the authors use tacit knowledge from domain experts about the events that might happen in the system, including contextual effects on the smart grid system. We argue that a context-aware explanation system may help lower such a dependency on tacit knowledge by learning the contextual influences on the behavior of CPS and continuously updating the context model with new information.

Explanation Visualization

To better enable human users to understand and reason about the behavior of CPS, researchers have in the past experimented with visualizing CPS' behavior by allowing users to observe the interaction between different components of a CPS through a "Magic Lens" using immersive technologies like augmented reality (Mayer, Hassan, and Sörös 2014). As a result, several new approaches are developing towards immersive visualization of the explanations (Frye, Rowat, and Feige 2020). Recent growth in adopting extended reality methods (for instance, on industrial shop floors) has opened up a range of possibilities to present explanations to the users. Some of the popular visualization techniques used in XAI for better user interpretation of the explanation are Tensor-flow graph visualization (Wongsuphasawat et al. 2017), Digital staining (Cruz-Roa et al. 2013), Limited Scoped Natural Language, and Heatmaps (Zeiler and Fergus 2014). However, the time-continuous nature of CPS is not well represented using the limited feature importance highlighting and overlaying techniques used in various explanation methods studied for this survey. Furthermore, (Schlegel et al. 2019) compares multiple explanation methods (LIME, SHAP, DeepLIFT) for incorporating temporal dimensions to conclude that most of the explanation methods work for specific architectures but are ineffective in conveying the result to increase users' understandability of CPS. (Schlegel et al. 2019) also points out the need for a more sophisticated visualization tool for time series explanations than overlaying of time series data in a heatmap. Such visualization techniques help users see the relationship and data exchange among the CPS and different entities in a context. However, because the interpretation responsibility is left to the user, there is a possibility of misunderstanding, which might result in the explanations becoming less understandable and actionable.

3 Discussion and Future Directions

The explanation techniques particular to CPS still require significant research work to make them interpretable and understandable for different types of users. Thus, we propose enhancements and recommend strategies that we believe would benefit the explanation of CPS' behavior:

The first enhancement that we propose to enable the application of XAI systems to CPS is what we refer to as *context-aware explanations*. We believe that a context-aware explanation system will assist users in comprehending the CPS' behavior in previously unconsidered instances (e.g., harsh working conditions and geographical relocation). One of the primary concerns in developing context-aware explanations using a data monitoring and analysis approach is to model the constantly evolving context of CPS. As discussed in Section 2, many recent studies use knowledge graphs to model CPS' context. However, those pre-built knowledge graphs should be updated to stay relevant in the dynamic context of CPS. A recent work that models the context of a pill dispenser system to explain its behavior (Sahlab, Jazdi, and Weyrich 2020) points out the uncertainty of such relationships due to their dynamic nature. Hence, we propose to use the counterfactual explanations method to learn the causal relationships among the factors influencing the behavior of CPS in dynamic contexts. Thus, freshly learned relationships and user feedback can be used to update the context model by the explanation system for further reasoning and future explanations.

In addition, an *enhanced explanation delivery mechanism* based on knowledge graph and counterfactual explanation methods might deliver explanations that are easier to comprehend. Because the explanation is based on the users' understanding of how the system works (in general) and the underlying relationships between physical properties. Users may have a better understanding of how the (cyber-physical) system works in general and why it behaves the way it does in a specific scenario when the CPS' context model and system knowledge are integrated with explanation visualization methodologies. Moreover, the intuitive presentation of explanations and additional information might increase the intelligibility of the explanations. Hence, these explanations can be actionable as users can then act upon the presented explanations using the updated knowledge graph and causal relationships, i.e., the users could adjust the CPS' context and input features to achieve the expected behavior.

4 Conclusion

In this paper, we proposed several extensions of explainability systems that would make XAI approaches amenable to explaining the behavior of CPS. We propose that integrating explanation approaches from XAI and semantic technologies, together with contextual information of the CPS, will enable the explanation systems to better explain the behavior of the CPS. Concretely, this could be accomplished through the outlined approaches of context-aware explanations and enhanced explanation learning and delivery.

References

Aryan, P. R.; Ekaputra, F. J.; Sabou, M.; Hauer, D.; Mosshammer, R.; Einfalt, A.; Miksa, T.; and Rauber, A. 2021. Explainable cyber-physical energy systems based on knowledge graph. https://doi.org/10.1145/3470481.3472704.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10.

Baheti, R., and Gill, H. 2011. Cyber-physical systems. *The impact of control technology* 12.

Blumreiter, M.; Greenyer, J.; Garcia, F. J. C.; Klös, V.; Schwammberger, M.; Sommer, C.; Vogelsang, A.; and Wortmann, A. 2019. Towards self-explainable cyber-physical systems. In *2019 ACM/IEEE 22nd International Confer-*

ence on Model Driven Engineering Languages and Systems Companion (MODELS-C).

Cruz-Roa, A. A.; Ovalle, J. E. A.; Madabhushi, A.; and Osorio, F. A. G. 2013. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Daglarli, E. 2021. Explainable artificial intelligence (xai) approaches and deep meta-learning models for cyber-physical systems. In *Artificial Intelligence Paradigms for Smart Cyber-Physical Systems*. IGI Global.

Doshi-Velez, F.; Kortz, M.; Budish, R.; Bavitz, C.; Gershman, S.; O'Brien, D.; Scott, K.; Schieber, S.; Waldo, J.; Weinberger, D.; et al. 2017. Accountability of ai under the law: The role of explanation.

Frye, C.; Rowat, C.; and Feige, I. 2020. Asymmetric shapley values: incorporating causal knowledge into model-agnostic explainability. *Advances in Neural Information Processing Systems* 33.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An approach to evaluating interpretability of machine learning.

Gusenbauer, M. 2019. Google scholar to overshadow them all? comparing the sizes of 12 academic search engines and bibliographic databases. *Scientometrics* 118.

Haque, S. A.; Aziz, S. M.; and Rahman, M. 2014. Review of cyber-physical system in healthcare. *international journal of distributed sensor networks* 10.

Kitchenham, B., and Charters, S. 2007. Guidelines for performing systematic literature reviews in software engineering.

Kopitar, L.; Cilar, L.; Kocbek, P.; and Stiglic, G. 2019. Local vs. global interpretability of machine learning models in type 2 diabetes mellitus screening. In *Artificial Intelligence in Medicine: Knowledge Representation and Transparent and Explainable Systems*.

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)*.

Lopez, F.; Saez, M.; Shao, Y.; Balta, E. C.; Moyne, J.; Mao, Z. M.; Barton, K.; and Tilbury, D. 2017. Categorization of anomalies in smart manufacturing systems to support the selection of detection mechanisms. https://doi.org/10.1109/LRA.2017.2714135.

Lundberg, S. M.; Erion, G.; Chen, H.; DeGrave, A.; Prutkin, J. M.; Nair, B.; Katz, R.; Himmelfarb, J.; Bansal, N.; and Lee, S.-I. 2020. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence* 2.

Mayer, S.; Hassan, Y. N.; and Sörös, G. 2014. A magic lens for revealing device interactions in smart environments. In *SIGGRAPH Asia 2014 Mobile Graphics and Interactive Applications*.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267.

Molnar, C. 2020. *Interpretable Machine Learning*.

Petnga, L., and Austin, M. 2013. Ontologies of time and time-based reasoning for mbse of cyber-physical systems. https://doi.org/10.1016/j.procs.2013.01.042.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "why should I trust you?": Explaining the predictions of any classifier.

Ricard, Q., and Owezarski, P. 2020. Ontology based anomaly detection for cellular vehicular communications.

Richens, J. G.; Lee, C. M.; and Johri, S. 2020. Improving the accuracy of medical diagnosis with causal machine learning. https://doi.org/10.1038/s41467-020-17419-7.

Sahlab, N.; Jazdi, N.; and Weyrich, M. 2020. Dynamic context modeling for cyber-physical systems applied to a pill dispenser. https://doi.org/10.1109/ETFA46521.2020.9211876.

Schlegel, U.; Arnout, H.; El-Assady, M.; Oelke, D.; and Keim, D. A. 2019. Towards a rigorous evaluation of xai methods on time series. https://doi.org/10.1109/ICCVW.2019.00516.

Shin, D. 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies* 146.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences.

Shrivastava, A.; Derler, P.; Baboud, Y.-S. L.; Stanton, K.; Khayatian, M.; Andrade, H. A.; Weiss, M.; Eidson, J.; and Chandhoke, S. 2016. Time in cyber-physical systems. In *Proceedings of the Eleventh IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks.

Weber, T., and Wermter, S. 2020. Integrating intrinsic and extrinsic explainability: The relevance of understanding neural networks for human-robot interaction.

Wickramasinghe, C. S.; Amarasinghe, K.; Marino, D. L.; Rieger, C.; and Manic, M. 2021. Explainable unsupervised machine learning for cyber-physical systems. *IEEE Access* 9.

Wongsuphasawat, K.; Smilkov, D.; Wexler, J.; Wilson, J.; Mane, D.; Fritz, D.; Krishnan, D.; Viégas, F. B.; and Wattenberg, M. 2017. Visualizing dataflow graphs of deep learning models in tensorflow. *IEEE transactions on visualization and computer graphics* 24.

Zeiler, M. D., and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*.