Poster: Towards Explaining the Effects of Contextual Influences on Cyber-Physical Systems

Sanjiv S. Jha
sanjiv.jha@unisg.ch
University of St.Gallen
St. Gallen, Switzerland

Simon Mayer
simon.mayer@unisg.ch
University of St.Gallen
St. Gallen, Switzerland

Kimberly García
kimberly.garcia@unisg.ch
University of St.Gallen
St. Gallen, Switzerland

ABSTRACT

The increasing complexity of Cyber-Physical Systems (CPS) increases the difficulty for users to understand their behavior. Using existing Explainable Artificial Intelligence (XAI) methods, CPS can explain their behavior to the users. However, the input-output correlations used in XAI methods are not capable of explaining certain anomalies on CPS behavior caused by contextual influences (CIs) since they do not consider the context of the CPS. Some well-known techniques used for understanding such CIs on CPS are test chambers and the analysis of logged CPS data. However, test chambers are typically only available to the manufacturer of a CPS, thus not useful for understanding CIs on the shop floors. Data analysis methods focus on data correlations, which are insufficient to explain causal relationships without using expert (human) knowledge. Hence, we propose a context-aware log-based explanation system to explain the causal relationship between CIs and the behavior of a CPS. The proposed solution employs semantic technologies to access the context of the CPS. It demonstrates the causal relationship between the CPS and CIs through counterfactual explanation and abductive reasoning methods. The contextual explanations offered by the proposed system will assist users in visualizing diverse scenarios in order to improve the CPS' behavior accordingly.

CCS CONCEPTS

• **Computing methodologies** → **Causal reasoning and diagnostics**; *Machine learning*; • **Human-centered computing** → *Text input*.

KEYWORDS

explainability, cyber-physical systems, knowledge graph, anomaly-detection, reasoning, context-aware

ACM Reference Format:
Sanjiv S. Jha, Simon Mayer, and Kimberly García. 2021. Poster: Towards Explaining the Effects of Contextual Influences on Cyber-Physical Systems. In *11th International Conference on the Internet of Things (IoT '21), November 8–12, 2021, St.Gallen, Switzerland.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3494322.3494359

IoT '21, November 8–12, 2021, St.Gallen, Switzerland

ACM ISBN 978-1-4503-8566-4/21/11…$15.00
https://doi.org/10.1145/3494322.3494359

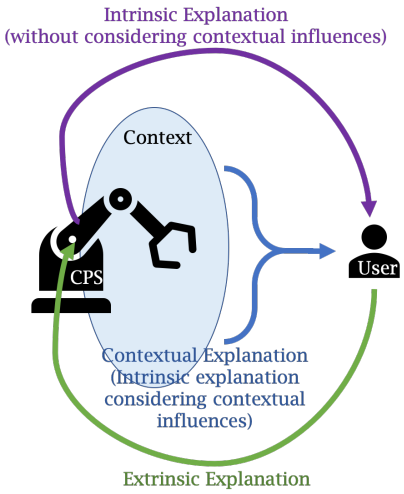

Figure 1: Contextual explanations provide additional information that is out of scope for common intrinsic explanation methods but relevant to CPS

1 INTRODUCTION

The advent of Industry 4.0 and the Internet of Things has made it possible for a range of physical systems used on shop floors to interact with the cyber space using algorithms. Cyber-Physical Systems (CPS) are employed to automate repetitive and complex processes for stakeholders (e.g., users, auditors, and engineers) using CPS functionalities like sensing and decisions-making. However, with the intricacy of these processes, the complexity of the algorithms used to achieve automation also increases. This lowers the behavioral transparency and trust of stakeholders in the system. Therefore, it is relevant for these CPS to explain their (black-box) behavior to stakeholders [2]. To this end, recent projects [18, 27] employ Explainable Artificial Intelligence (XAI) methods such as Local Interpretable Model-agnostic Explanations (LIME) [19], Shapley Additive Explanations (SHAP) [13], or Counterfactual Explanation [14] for system-generated behavioral explanations. According to [27], explanations can be categorized as *intrinsic* and *extrinsic*. Explanations are *intrinsic* when offered by the machines to the users, and *extrinsic* when fed to the CPS from some external entities. Weber and Wermter [27] claim that XAI research (including LIME, SHAP, and the example-based explanations using counterfactuals) has focused so far on producing intrinsic explanations. Intrinsic explanations can explain a system's behavior on local and global scopes. LIME's objective is to explain a particular decision made by

a system, and SHAP describes the overall behavior of a system [14]. However, even if the current intrinsic explanation algorithms could explain most of the regular decisions CPS make, they do not consider the effects of the (dynamic) contexts in which most CPS are deployed. Therefore, many anomalies caused by contextual factors might remain unresolved, and unexplained. Hence,we argue that for explanations to be relevant in contextually dynamic scenarios, explanation systems should stop considering CPS as isolated entities. Instead, such systems need to take into account contextual factors (e.g., location, time, and environmental parameters) that might affect the behavior of the CPS in addition to variables that describe its internal state. We refer to Intrinsic explanations considering the contextual factors as *Contextual Explanations*. Figure 1 illustrates the proposed contextual explanations with respect to intrinsic and extrinsic explanations.

It is natural for users to anticipate CPS to behave in expected ways in response to known inputs. However, the effect of contextual factors on a CPS can cause anomalies in its behavior. In this paper, we define *Contextual Influence* (CI) as *changes of the values of contextual parameters that could potentially affect the behavior of a CPS*. Such CIs might have short- and long-term consequences, such as porosity issues in additive manufacturing due to frequently changing ambient temperature [8] or corrosion of electronics due to a high level of ambient humidity on the shop floor for an extended period – in these examples, the CIs are the ambient temperature and ambient humidity, respectively. To investigate CIs of a CPS, it is essential to first define the *context* of this CPS. Following different definitions of context coined over the past decades, such as Dey's popular definition [7] and a recent CPS-oriented one [22], we define the system-oriented context of a CPS as *the observable entities in the physical or virtual vicinity of a CPS that could influence the functioning of such a CPS*. We provide further details on context and its modeling for the proposed system in Section 3.

The following scenario illustrates the concept of CI in this paper: A user keeps their 3D printer in *Room 1* near the window during weekdays and then moves it to *Room 2* in the basement (with no windows) on the weekends. The nozzle of this 3D printer for PLA printing typically heats up at $210^{\circ}C$ in 3 minutes. However, the user observes that the nozzle heats up faster in *Room 2* than in *Room 1*, i.e., it requires only 2 minutes in *Room 2*. This behavior might raise questions regarding this anomalous (temporal or spatial) behavior of the CPS. For example, a user query can be "Why does PrinterMKS3 take longer to warm up?". The reasoning, in this case, relies on two contextual factors: time and location of the machine. Here the time acts as a reference axis on which one would substantiate the causability: *"the extent to which an explanation of a statement to a human expert achieves a specified level of causal understanding with effectiveness, efficiency, and satisfaction in a specified context of use"* [10]. Causal reasoning using the context information may then help determine that the cold air flowing in through the open window in *Room 1* is the cause for the 3D printer's behavioral change. Moreover, the causal inference (through abductive reasoning [6]) could help understand a relationship between the windows and the CI (i.e., the decrease in ambient temperature because of the cold air flowing in through the open window) on the ambient room temperature that influences the nozzle's initial temperature.

In this paper, we propose a methodology for creating an explanation system that explains the behavior and behavioral anomalies of CPS caused by CIs. The system prototype is underway, thus not presented in this article.

2 RELATED WORK

One of the popular tools used in the industries to understand the CI on a machine in foreign conditions is commonly known as a *test chamber*. Test chamber applications such as indoor air chemistry and electronic office equipment testing are described in [23]. Test chambers emulate different contextual scenarios (e.g., extreme heat or high humidity) around a machine to observe any changes on the machine's behavior. Physical test chambers come in various shapes and sizes to accommodate the needs of stakeholders. However, their customization and maintenance are expensive and complex. Hence, test chambers are typically used by large CPS manufacturers, but are much less used by small and medium-sized enterprises (SME). They are often not accessible to the customers that operate CPS on their premises, which means that today, the monitoring of CI is not practical in *deployed* CPS but constrained to pre-shipment testing. Consequently, a cost-effective and scalable system for SME deployment sites to explain the causal relationship between CI and the observed CPS behavior is required.

Apart from test chambers, monitoring and analysing methods are also used to understand behavioural anomalies in CPS. Recent studies have introduced several techniques for detecting and reasoning about behavioral anomalies in CPS via various model-based and model-free mechanisms like feature extraction with limit checking, clustering, classification, and Knowledge-based methods [1, 12]. Since CPS are composed of many components, monitoring them generates a large amount of log data. This leads to high data velocity and dimensionality issues when analyzing the resulting (typically heterogeneous) data. One of the solutions to this issue is to lower the dimensionality of the data using methods like PCA [26] and control its volume using contextual information. Hence, there are some initiatives for decreasing the training data size through context-awareness in machine learning approaches, that is, selecting training data based on the context (i.e., season of the year) as proposed in [16]. However, while explaining the behavior of CPS that are using such machine learning models, the generic intrinsic explanations fall short in considering CI.

Current XAI methods do not offer simple and understandable explanations for time series data [24]. Since CPS exhibit time-continuous behavior through their log data, timestamps are preferred to detect instantaneous or event-based anomalies in CPS. Several methods proposed in recent work [17, 20] consider time and knowledge-based reasoning to detect and predict anomalies. Hence, there is a need to create an explanation system that is able to use time-series data for generating explanations. One recent study [17] uses time synchronization of two entities (yellow light traffic signal and a smart car's decision-making module) to reason the decisions made by the smart car for safety and risk assessment using ontologies. Clock synchronization [9, 25] between contextual and input variables of the CPS is thus required to identify possible unexplained causes for anomalies in the CPS behavior by cross-correlation of gathered data. Moreover, an existing example-based

method in XAI – counterfactual explanation [14] – can also provide a similar causal relationships without the needed expert knowledge to build a Knowledge Graph (KG). The prevalent counterfactual explanation techniques [3, 15, 28] in XAI are promising methods for delivering correct interpretations and relationships between the input values and model's decisions [21]. However, only some of the counterfactual explanation methods are user-tested [11]. Moreover, to understand the CI on CPS, it is crucial to access the context data. To this end, several projects [17, 22] use semantic technologies to model and scope the context of a CPS.

In this work, we propose a contextual explanation system that utilizes time-synchronized time-series log data produced by a CPS and its context to explain its behavior. Unlike the test chambers and prevalent anomaly detection methods, the proposed system presents a cost-effective and scalable solution for contextual explanations. The relationships learned through counterfactuals in a new context scenario will be used to enrich a KG that could be used for future causal relationships and explanation generation.

3 A CONTEXTUAL EXPLANATION SYSTEM FOR CPS

To scope and access the observable context of a CPS we use semantic technologies, since an ontology provides an interoperability layer capable of aggregating data from different information sources.

3.1 Context Modeling

The contextual explanation system proposed in this paper uses a Context Discovery component, similar to the crawler component described in [5]. Such a component allows the proposed system to find entities (e.g., sensors) capable of providing contextual information that describes CI on the CPS. This Context Discovery component crawls a hypermedia environment to find machine-readable metadata that describes the application programming interfaces of sensors and actuators present in that context in the form of W3C WoT Thing Descriptions (TD)[1]. Thus, from a query typed by a user, keywords referring to a CPS are obtained to retrieve the CPS' TD as well as TDs of devices that are related to the CPS. For example, if a user makes a query containing the CPS name "PrinterMKS3", the Context Discovery component checks the location of the printer using its TD and lists all the found TDs in that location.

Moreover, the Context Discovery component interacts with a KG that has been built with a skeleton of a CPS, but which is automatically enriched as causal relationships among the entities in the CPS and CI are learned (using counterfactuals and abductive reasoning). Since the context of CPS is dynamic and constantly evolving, the pre-built KG used to model its context needs to be updated accordingly to stay relevant in the changing scenarios. Traditionally, in semantic technologies updating the schema of a KG was dependent on expert knowledge. However, in recent years many automated techniques have been proposed to enrich a KG with new knowledge [4]. A recent work [22] uses KG to model context, and its relationship with the CPS highlights that such relationships are uncertain due to its dynamicity. Thus, supporting methods for discovering relationships should be considered. Hence, to provide better causal reasoning and to lower the dependency of

[1]https://www.w3.org/TR/wot-thing-description/

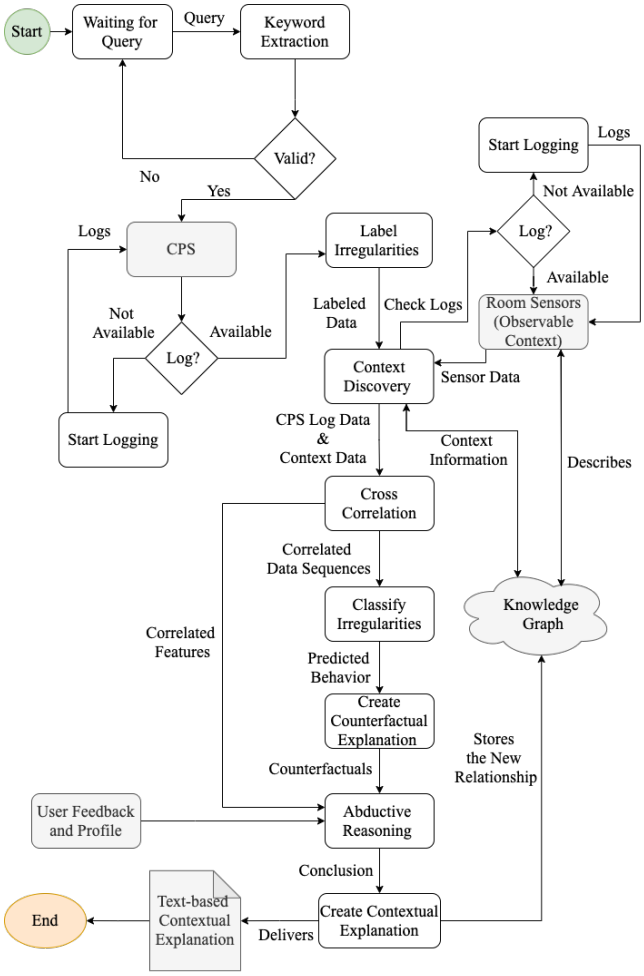

Figure 2: Based on queries submitted by the users, the system generates contextual explanations with causability

the explanation system on expert knowledge, our proposed solution employs counterfactual explanations and an abductive reasoning technique (similar to the logical abduction method described in [6]) for creating possible explanations to learned relationships. Therefore, the system described in this paper could help deliver an initial explanation for anomalies that are caused by the CI of a CPS.

3.2 System Data Flow

The operational flow of the proposed system is illustrated in Figure 2. It consists of the following steps:

- **1.1 Keyword Extraction:** User submits a query that is parsed by the system to extract keywords (e.g., PrinterMKS3, longer, and warm are the keywords in our example query "Why does PrinterMKS3 take longer to warm up?"). Once the keywords are detected the system treats this query as a valid query and moves forward.
- **1.2 Log Collection:** The system then asks the CPS, if the logs for extracted features are already available. If logs are available, the system moves to step 2; otherwise, the user triggers the

devices for log collection over a specific amount of time, which results in a delayed explanation.

2 **Label Irregularities:** The log is checked and labeled by the system for irregularities in the data corresponding to the extracted keywords from the user's query.

3 **Context Discovery:** Depending on the extracted keywords, the Context Discovery component lists all the entities available in the context. Each of the discovered entities provides observed data (if logged data is available, otherwise step 1.2 is followed again) that are aligned sequentially using their timestamps.

4 **Cross Correlation:** The context data and CPS log data are cross-correlated to identify highly correlated time-series data sequences and other contextual factors are discarded.

5 **Classify Irregularities:** The system feeds the labeled dataset to a supervised machine learning classification model to predict the anomaly.

6 **Create Counterfactual Explanation:** The Counterfactual Explainer component (such as the counterfactual explanation methods described in [3, 15]) takes the predicted behavior of the CPS and creates possible counterfactuals for the causal relationship between relevant contextual factors and the CPS. Hence, the user can see the suggested amendments in the correlated contextual factors to switch the behavior of the CPS (predicted by the classification model) from the anomalous to the regular (class).

7 **Abductive Reasoning:** Further the abductive logic program (similar to one described in [6]) concludes the causation through user-selected counterfactuals and context information (used as hypothesis and fact base).

8 **Create Contextual Explanation:** The system synthesizes a comprehensive text-based explanation (Contextual Explanation) using generated counterfactuals and the abductive conclusions. At this step, the users understand the relationship between the PrinterMKS3 and the change in ambient temperature. Further, the learned causal relationship, CPS's name/identifier, extracted keywords (from user's question), and the textual explanation are added to the KG by creating relevant relationships between the CPS and machine-readable descriptions of the contextual entities. Such KG could be used to learn relationships to generate possible contextual explanations in the future.

4 CONCLUSION

In this article, we have proposed a solution for understanding CIs on CPS, which can be used to detect and quantify contextual influences on a CPS. Using this solution, we hypothesize that stakeholders will be able to not only better understand such influences, but also to optimize the behavior of a CPS by maintaining a favorable context around it. The article describes the concept of such a Contextual Explanation system. We are currently working on a first proof of concept implementation to assess the generated contextual explanations (delivered in textual form) by the proposed solution in various contextual scenarios. As part of the future work, a qualitative study will be conducted to validate the causability and understandability of the explanations by stakeholders.

REFERENCES

[1] Marcello Balduccini, Edward Griffor, Michael Huth, Claire Vishik, Martin Burns, and David Wollman. 2018. Ontology-based reasoning about the trustworthiness of cyber-physical systems. (2018).

[2] Mathias Blumreiter, Joel Greenyer, Francisco Javier Chiyah Garcia, Verena Klös, Maike Schwammberger, Christoph Sommer, Andreas Vogelsang, and Andreas Wortmann. 2019. Towards self-explainable cyber-physical systems.

[3] Dieter Brughmans and David Martens. 2021. NICE: An Algorithm for Nearest Instance Counterfactual Explanations. (2021).

[4] Xiaojun Chen, Shengbin Jia, and Yang Xiang. 2020. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications* 141 (2020).

[5] Andrei Ciortea, Simon Mayer, Simon Bienz, Fabien Gandon, and Olivier Corby. 2020. Autonomous search in a social and ubiquitous Web. *Personal and Ubiquitous Computing* (2020).

[6] Wang-Zhou Dai, Qiuling Xu, Yang Yu, and Zhi-Hua Zhou. 2019. Bridging machine learning and logical reasoning by abductive learning. (2019).

[7] Anind K Dey. 2001. Understanding and using context. *Personal and ubiquitous computing* 5 (2001).

[8] Lichen Fang, Yishu Yan, Ojaswi Agarwal, Shengyu Yao, Jonathan E Seppala, and Sung Hoon Kang. 2020. Effects of Environmental Temperature and Humidity on the Geometry and Strength of Polycarbonate Specimens Prepared by Fused Filament Fabrication. *Materials* 13 (2020).

[9] Jerry Fowler and Willy Zwaenepoel. 1990. *Causal distributed breakpoints.* Technical Report.

[10] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9 (2019).

[11] Mark T Keane, Eoin M Kenny, Eoin Delaney, and Barry Smyth. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. (2021).

[12] Felipe Lopez, Miguel Saez, Yuru Shao, Efe C Balta, James Moyne, Z Morley Mao, Kira Barton, and Dawn Tilbury. 2017. Categorization of anomalies in smart manufacturing systems to support the selection of detection mechanisms. *IEEE Robotics and Automation Letters* 2 (2017).

[13] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.).

[14] Christoph Molnar. 2019. *Interpretable Machine Learning*.

[15] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*.

[16] Nathalia Nascimento, Paulo Alencar, Carlos Lucena, and Donald Cowan. 2018. A context-aware machine learning-based approach.

[17] Leonard Petnga and Mark Austin. 2013. Ontologies of time and time-based reasoning for MBSE of cyber-physical systems. *Procedia Computer Science* 16 (2013).

[18] Jana-Rebecca Rehse, Nijat Mehdiyev, and Peter Fettke. 2019. Towards explainable process predictions for industry 4.0 in the dfki-smart-lego-factory. *KI-Künstliche Intelligenz* 33 (2019).

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016.*

[20] Quentin Ricard and Philippe Owezarski. 2020. Ontology Based Anomaly Detection for Cellular Vehicular Communications. In *10th European Congress on Embedded Real Time Software and Systems (ERTS 2020).*

[21] Jonathan G Richens, Ciarán M Lee, and Saurabh Johri. 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature communications* 11 (2020).

[22] Nada Sahlab, Nasser Jazdi, and Michael Weyrich. 2020. Dynamic Context Modeling for Cyber-Physical Systems Applied to a Pill Dispenser.

[23] Tunga Salthammer. 2009. Environmental test chambers and cells. *Organic indoor air pollutants: occurrence, measurement, evaluation* (2009).

[24] Udo Schlegel, Hiba Arnout, Mennatallah El-Assady, Daniela Oelke, and Daniel A Keim. 2019. Towards a rigorous evaluation of XAI Methods on Time Series. (2019).

[25] Mukesh Singhal and Ajay Kshemkalyani. 1992. An efficient implementation of vector clocks. *Inform. Process. Lett.* 43 (1992).

[26] Srikanth Thudumu, Philip Branch, Jiong Jin, and Jugdutt Jack Singh. 2020. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data* 7 (2020).

[27] Tom Weber and Stefan Wermter. 2020. Integrating Intrinsic and Extrinsic Explainability: The Relevance of Understanding Neural Networks for Human-Robot Interaction. (2020).

[28] Adam White and Artur d'Avila Garcez. 2019. Measurable counterfactual local explanations for any classifier. (2019).