Knowledge Graph for Explainable Cyber Physical Systems: A Case study in Smart Energy Grids*

Peb Ruswono Aryan[1][0000-0002-1698-1064]

Vienna University of Technology, Vienna, Austria
peb.aryan@tuwien.ac.at

Abstract. The rapid development of computing technology and automation widens the scope of the task delegated to cyber-physical systems (CPS) such as smart grids or smart buildings. Explainability, i.e., the ability to provide explanations about system states or behaviors becomes one of the requirements for future cyber-physical systems as more complex computer-made decisions affect our daily lives. The work on the explainability in CPS is scarce despite recent attention on the explainability of algorithms in artificial intelligence. This doctorate research aims to comprehensively understand the scope of explainability in CPS, identify the critical components of an explainable CPS, and methods and metrics to evaluate them. Specifically, our main research question is how and to what extent Knowledge Graphs can be applied in enabling the explainability of CPS. Using the design science approach, we attempt to answer these questions in a set of iterations, starting with a simulation-based approach and constructing a baseline system followed by more focused studies and more realistic settings using data from real-world CPS. The selected application domain in this work is industrial energy systems such as smart grids and smart buildings. The expected outcome of this work is a theoretical foundation and methods for developing an explainable CPS applicable in various domains.

Keywords: knowledge graph · explainability · cyber-physical systems.

1 Problem Statement

Recently there has been concern that future CPS which span both the realm of physical and cyber-worlds are challenged to explain their behavior to users, engineers, and other stakeholders [7]. Rapid technological development in the digital aspect of CPS, such as communication, control, and computation, drives the increasing scale and complexity of CPSs. For example, low-power wireless communication allows the proliferation of objects connected through a more extensive network. Advances in machine learning allow data processing algorithms

* Supported by Siemens AG and Austrian Research Promotion Agency (FFG) in the PoSyCo project (FFG No. 3036508).

to become adaptive and capable of solving complex real-world tasks. When these complexities gain more influence on systems that impact our day-to-day life, the necessity of having explanations in terms of the behavior of systems is emerging.

Explainability is the ability of a (software) system to provide explanations about its states or behaviors in terms of a set of facts [14]. Explanations foster understanding of a thing being explained (explanandum) by linking it with existing knowledge on the receiving stakeholder's side [11]. Recent studies about explainability are primarily oriented towards artificial intelligence (AI)-based methods which function as black-boxes, meaning that their decisions are not transparent to end users [3]. Explainability is also an emerging issue in more complex systems, such as cyber-physical systems. However, limited research has been performed so far on understanding the theoretical foundations of explainability in CPS as well as on exploring suitable solution paradigms to this problem.

Exploring various risk-related scenarios is undesirable to be conducted in the real system, that is *in vivo*. Having a *in vitro* platform and reusable framework allows for a more rapid development process and avoiding the unnecessary cost of trial and error when developing an explainable CPS.

As an illustrative example in the energy domain, smart electricity grids evolve from static to dynamically changing networks of large numbers of devices, e.g., photo-voltaic units (PV), electric vehicle charging stations (EVCS). The slow charging of an EVCS is an event that requires an explanation for several stakeholders including the EVCS owner, customer service representatives, field engineers, and grid planners. An explanation could be that overcast weather leads to lower than usual energy production through PVs in the region, this leads to a lack of supply in the grid segment and to a control intervention to reduce charging power by the grid operator. From the consumer's perspective, the change in energy consumption should be seen as independent of how it is produced. The energy production is then expected to run uninterrupted and provide sufficient supply even when there is an increase in consumption. A swift response is desired if an unwanted event such as failure to fulfill expected service or blackout is unavoidable since the loss caused by the fault would be a function of time. On the one hand, the technical operation employees expect detailed explanations in order to be able to decide the next course of action to remedy a potential fault/anomaly. On the other hand, the possibly larger population of affected consumers, an ideal explanation would be more succinct and related to their context, such as the service contract.

In order to generate perspicuous explanations for the intended stakeholders, the explanatory system needs to integrate information from various sources such as the structure of the system, the relationship between elements of the system, and the history of the system's state. Additionally, understanding the recipient of an explanation is also essential to be tailored to be easy to understand.

2 Related work

Artificial intelligence, mainly the area of expert and knowledge-based systems, extensively studies the task of providing explanation based on formalized logical reasoning [14]. Recently the necessity to provide explainable reasoning for the complex network has been reignited due to rapid progress in practical applications of machine learning in particular deep architectures of artificial neural networks. Explainability becomes a hot topic in the AI community following the concern about the ethical implications of applying machine learning solutions under biased data [3]. Explainable AI techniques developed to explain complex machine learning models to the users suggest that user orientation as one of the critical aspect of explainability[10].

The interpretation of explainability from the perspective of industrial systems is even more pragmatic. Related topics such as anomaly detection and subsequent root-cause analysis are essential topics in industrial (cyber-physical) systems and are currently achieved with methods such as FMEA (Failure Mode and Effect Analysis) [4] and FTA (Fault Tree Analysis) [6]. These methods require the specification of possible anomalies and their causes by various experts that know (parts of) the system and are typically hampered by the ambiguity and inconsistency of the collectively collected knowledge. The inconsistent terminology also hampers deriving meaningful explanations as a follow-up step of identifying a root cause for a given defect. Because of its specification in natural language, FMEA knowledge is difficult to reuse, is incomplete, and likely inconsistent (as there no formal way to check consistency) [5].

CPS, particularly the smart grid, is relatively new and evolving, and it combines different disciplines such as physics, statistics, and socio-economics. Studies of explainability in CPS are scarce, especially for specific topics such as the approaches based on the knowledge graph. One of the closest approaches is fault diagnosis systems in a smart building that combines a physical process model and data-driven approach[13]. This work builds causality knowledge from experts into a knowledge graph and applies SPARQL update rules to infer potential causes of a given event. Considering the multi-disciplinary nature of CPS, different communities use different representations of causality knowledge for solving different tasks. For example, in the community of distributed systems and cloud computing, one tries to automate causality mining from time-series data using correlation[15]. In the other community, i.e., energy and power systems, an ensemble of statistical causal models and deep neural network[16] are used to build models for short-term forecasting.

In summary, explainability encompasses, on the one end, a human who needs an explanation and, on the other end, causality knowledge that is not known explicitly from the system's description. Existing literature addressed these issues only partially, and the focus of different communities is diverse. Only by collecting various puzzle pieces can we see the big picture and establish a solid foundation of explainable CPS.

3 Research Questions

The literature suggests that there is a limited understanding of what constitutes an explainable cyber-physical systems. One line of research focuses on only user aspect of explanation while other line struggles with ad-hoc or partial solutions. Therefore, the main question for this research is:

RQ0 What are the main theoretical, methodological, and engineering foundations that enable effective and efficient implementation of explainable CPS?

This question is the starting point to the more specific questions: What are the core components needed to achieve explainability? To what extent knowledge graph can help build an explainable CPS? What are the requirements for making an explainable system applicable to various domains?

The literature also indicates that causality knowledge and user-oriented explanation generation algorithms are the critical aspect of an explainable CPS. Thus, this research also aims to address the following questions:

RQ1 *What is the effective semantic representation to integrate different representations of causality knowledge?* A different source of causality knowledge may have a different meaning of weight of a causal relationship. Some may involve the coefficient of a differential equation, and some others might refer to probabilistic quantities to refer to subjective belief or derived quality metrics.

RQ2 *How to acquire causality knowledge efficiently from data and domain experts?* How can we support domain experts to express causal relationships based on domain knowledge and data? One approach to express causality captured from domain experts used in [2] is SPARQL. How can we aid domain experts to express their knowledge without learning about SPARQL first? Can we acquire causality knowledge by analyzing temporal data using time-series analysis or machine learning?

RQ3 *What are effective and efficient algorithms for generation and ranking of (alternative) explanations?* What are the criteria to decide that an explanation is plausible? Given that there are multiple competing hypotheses, what metrics can be applied to compare and rank multiple explanation alternatives?

RQ4 How to effectively present explanations to system end-users? What are the cognitive aspects of a user that are important in determining whether an explanation is understandable or not? What are the metrics used for measuring the comprehensibility of an explanation output on a selected user model?

The questions above correspond to the core functionalities of explanation generation: causality knowledge acquisition and exploitation. Other aspects of

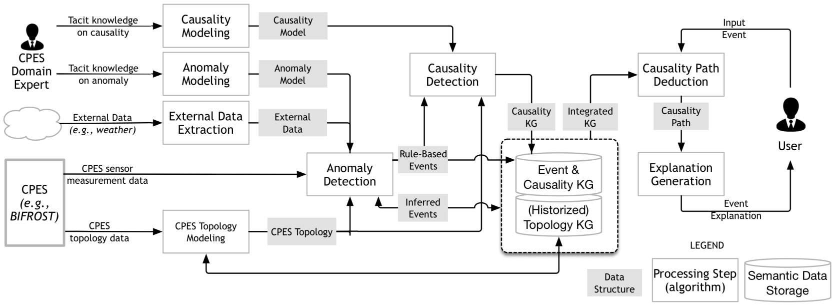


Fig. 1. Architecture of explainable CPS built on smart grid simulation platform [1, 2]

the interface to the end-users, such as visualization or generation of explanation in natural language, will not be addressed in this research. These aspects will become more apparent when the core and realistic use case scenarios have been developed.

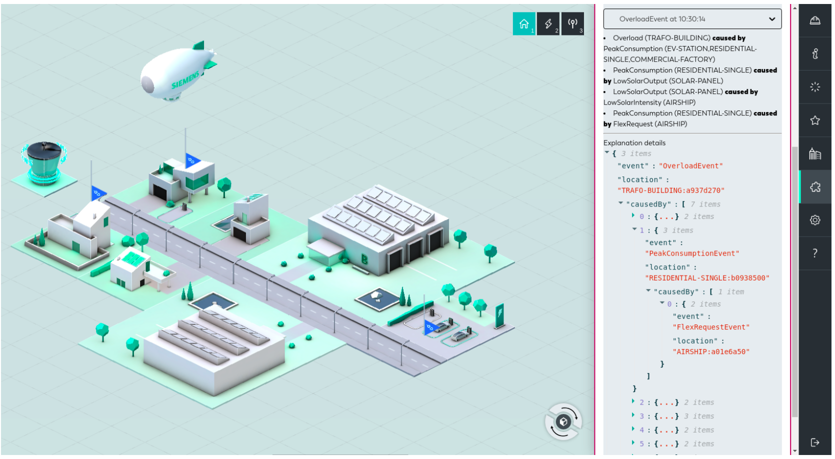
4 Research Plan and Preliminary results

This study adopts a classical Design Science methodology [8]: (i) the *rigor cycle* is ensured by grounding methods for answering all research questions from a thorough understanding of relevant literature studies and dissemination of the intermediate results in the scientific community; (ii) deriving requirements from concrete application contexts using simulation and living lab data from ongoing research projects and the creation of PoCs to address these requirements constitute the *relevance cycle*; (iii) method development, testing, and subsequent revision constitute the *design cycle*.

This study has been conducted for a year. The following 2-3 years will be focused on answering each research question. Specifically, the second year will be allocated for addressing the representation and acquisition (RQ1) of causality knowledge (RQ2). The analytics (RQ3) and presentation (RQ4) aspect of the explanation will be conducted in the third year.

Preliminary results The current state of the PhD has resulted in an understanding of explainability and explainable CPS. In the first iteration, the work is oriented towards understanding the explainability in energy systems. Developing plausible and feasible scenarios and data acquisition drives the focus on using a simulation platform (i.e. BIFROST [12], see Fig. 2). Additionally, the general idea of an explanation generation algorithm was developed and evaluated using synthetic data of a scenario related to electric car charging [1].

The following iteration builds upon the previous idea with more concrete artifacts. One of the results is the architecture shown in Figure 1. The realization of this architecture was then implemented as a prototype application for demonstrating that the explanations from the scenario can be derived based on



OverloadEvent at 10:30:14

- Overload (TRAFO-BUILDING) **caused by** PeakConsumption (EV-STATION,RESIDENTIAL-SINGLE,COMMERCIAL-FACTORY)
- PeakConsumption (RESIDENTIAL-SINGLE) **caused by** LowSolarOutput (SOLAR-PANEL)
- LowSolarOutput (SOLAR-PANEL) **caused by** LowSolarIntensity (AIRSHIP)
- PeakConsumption (RESIDENTIAL-SINGLE) **caused by** FlexRequest (AIRSHIP)

Explanation details

```
{ 3 items
  "event" : "OverloadEvent"
  "location" :
  "TRAFO-BUILDING:a937d270"
  "causedBy" : [ 7 items
    0 : {...} 2 items
    1 : { 3 items
      "event" :
      "PeakConsumptionEvent"
      "location" :
      "RESIDENTIAL-SINGLE:b0938500"
      "causedBy" : [ 1 item
        0 : { 2 items
          "event" :
          "FlexRequestEvent"
          "location" :
          "AIRSHIP:a01e6a50"
        }
      ]
    }
    2 : {...} 2 items
    3 : {...} 3 items
    4 : {...} 2 items
    5 : {...} 2 items
```

Fig. 2. Prototype of explainable CPS built on smart grid simulation platform

simulated data and captured knowledge. The solution design and implementation result was then published in the energy community proposing the solution based on semantic web technologies [2]. To this end, an ontology [1] for modeling data and knowledge described in the architecture has been developed.

Figure 2 displays the prototype of an explainable CPS build as a part of the BIFROST smart grid simulation engine.

Behind the user-facing interface is the engine that integrates data coming from the simulation into knowledge graphs in the triple store, deducing causal relations, detecting events, and deriving explanation for the detected events. The explanation is then displayed as shown on the right side of the figure. From this first iteration, we better understand what aspects are needed to build an explainable CPS.

5 Expected results and their evaluation

The goal of an explainable CPS is to provide explanations of events for a variety of scenarios. Close collaboration with domain experts is necessary To achieve plausible scenarios that can be used as a basis for further evaluation. The developed scenarios are then implemented in a simulation to generate data. Furthermore, actual measurements will be used to ensure the validity of the simulation data. User studies and empirical analysis are performed to evaluate the research questions. The following describes the outcome and outputs for each research question.

RQ1 The outcome for **RQ1** is the incorporation of different causality representations to generate an explanation. A vocabulary for different meanings (e.g.,

[1] https://pebbie.org/expcps/

relationship weight) of causality will be designed to augment the basic model of causality. Additionally, an algorithm to fuse these different semantics of causality will be developed as part of the explanation generation algorithm.

RQ2 Answering RQ2 will involve user studies and implementation of algorithms to derive causality knowledge from simulated data. Method to acquire causality knowledge will be developed and evaluated using user-study and empirical analysis.

RQ3 A set of metrics and ranking algorithms will be developed, and an explanation generation task will be executed based on a prepared scenario. A group of domain experts will be asked to manually create explanations as the gold standard for measuring the algorithm's performance. The evaluation of the algorithm will use metrics such as MRR or Precision@10 as a base and modified to accommodate comparing the graph structure of the explanation.

RQ4 A qualitative study will be conducted to acquire key characteristics of explanation target or user profiles. A set of user-profiles will be defined, and the explanation generation task will be executed using the scenarios. Another user study will assess whether the customized generated explanation is relevant to the intended explanation recipients. To this end, System Causability Scale (SCS) [9] will be used as one of the metrics.

6 Discussion and Future work

The previous section described an end-to-end prototype of how an explainable CPS should work after the first year of the doctoral study. This initial work helps to identify issues formulated in the research questions for the next iteration. The collection of artifacts from investigating each research question forms a solution framework to build an explainable CPS.

Some topics possibly related to explainable CPS are intentionally not addressed considering the limited scope of this research. e.g., such as scalable storage techniques to handle large-scale CPS data. Other topics depend on the mentioned research questions, such as the study of specific presentation forms of explanation (e.g., visualization) or exploratory search systems to explore the explanation hypothesis and history of events. Further research on these topics will enrich the framework for building an explainable CPS and enables explainability in various systems.

7 Acknowledgement

This work is funded through a grant project titled "Power System Cognification" (PoSyCo) by the Austrian Research Promotion Agency (FFG) with Siemens AG. I am thankful for PhD supervisory team Marta Sabou, Fajar Juang Ekaputra, and Tomasz Miksa for the professional advises and personal supports.

References

1. Aryan, P.R., Ekaputra, F.J., Sabou, M., Hauer, D., Mosshammer, R., Einhalt, A., Miksa, T., Rauber, A.: Simulation support for explainable cyber-physical energy systems. In: 8th Workshop on Modeling and Simulation for Cyber-Physical Energy Systems (MSCPES2020) (2020)
2. Aryan, P.R., Ekaputra, F.J., Sabou, M., Hauer, D., Mosshammer, R., Einhalt, A., Miksa, T., Rauber, A.: Explainable cyber-physical energy systems based on knowledge graph. In: 9th Workshop on Modeling and Simulation for Cyber-Physical Energy Systems (MSCPES2021) (2021)
3. Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., et al.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (Jun 2020). https://doi.org/10.1016/j.inffus.2019.12.012
4. Ben-Daya, M.: Failure Mode and Effect Analysis, pp. 75–90. Springer London, London (2009)
5. Dittmann, L., Rademacher, T.T., Zelewski, S.: Performing FMEA Using Ontologies. In: 18th International Workshop on Qualitative Reasoning. pp. 209–216 (2004)
6. Ericsson, C.A.: Fault Tree Analysis Primer (2011)
7. Greenyer, J., Lochau, M., Vogel, T.: Explainable Software for Cyber-Physical Systems (ES4CPS): Report from the GI Dagstuhl Seminar 19023, January 06-11 2019, Schloss Dagstuhl (2019)
8. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design science in information systems research. MIS quarterly pp. 75–105 (2004)
9. Holzinger, A., Carrington, A., Müller, H.: Measuring the quality of explanations: the system causability scale (scs). KI-Künstliche Intelligenz pp. 1–6 (2020)
10. Kirsch, A.: Explain to whom? putting the user in the center of explainable ai. In: Proceedings of the First International Workshop on Comprehensibility and Explanation in AI and ML 2017 co-located with 16th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2017) (2017)
11. Lombrozo, T.: The structure and function of explanations. Trends in cognitive sciences **10**(10), 464–470 (2006)
12. Mosshammer, R., Diwold, K., Einfalt, A., Schwarz, J., Zehrfeldt, B.: BIFROST: A Smart City Planning and Simulation Tool. In: Karwowski, W., Ahram, T. (eds.) Intelligent Human Systems Integration. pp. 217–222. Springer (2019)
13. Ploennigs, J., Schumann, A., Lécué, F.: Adapting Semantic Sensor Networks for Smart Building Diagnosis. In: 13th International Semantic Web Conference (ISWC). pp. 308–323. Springer (2014). https://doi.org/10.1007/978-3-319-11915-1_20
14. Preece, A.: Asking 'Why'in AI: Explainability of intelligent systems–perspectives and challenges. Intelligent Systems in Accounting, Finance and Management **25**(2), 63–72 (2018)
15. Qiu, J., Du, Q., Yin, K., Zhang, S.L., Qian, C.: A causality mining and knowledge graph based method of root cause diagnosis for performance anomaly in cloud applications. Applied Sciences **10**(6), 2166 (2020)
16. Sriram, L.M.K., Ulak, M.B., Ozguven, E.E., Arghandeh, R.: Multi-network vulnerability causal model for infrastructure co-resilience. IEEE Access **7**, 35344–35358 (2019)