Semantics-Enabled System Transparency: User-Centered Explanations in Cyber-Physical Systems

Katrin Schreiberhuber[1]

[1]*Wirtschaftsuniversität Wien, Welthandelsplatz 1, 1020 Vienna, Austria*

Abstract

A Cyber Physical System (CPS) helps tackle complex problems in various domains, e.g., smart grids or smart buildings. As the complexity of such systems' behaviours often reduces user understanding and trust, this research proposal aims to enhance system transparency and user-centered explanations in such CPSs. Despite the increasing adoption of CPS across various domains, existing solutions for explaining system behaviors remain limited in scope and effectiveness. We propose a domain-independent framework, leveraging Semantic Web Technologies (SWT), to tackle this challenge. The framework aims to bridge the gap between system complexity and user comprehension by providing user-centered explanations tailored to different user types, advancing transparency, understanding, and usability in CPS deployments. Furthermore, by extending the application domain of SWT to Explainable Cyber Physical System (ExpCPS), this research contributes to the broader semantic web community. Preliminary results have shown the feasibility of the approach in a small use case, based on the newly developed ExpCPS ontology and a rule-based explanation module to derive causal paths from explicit knowledge. Future work will focus on extending these solutions as well as research on the design and evaluation of user-centered explanations.

Keywords

Cyber-Physical System, Explainability, user-centered Explanations, Causality Representation

1. Context and Motivation

A CPS represents the integration of computational and physical components to enhance the intelligence, efficiency and handling of smart systems and their processes. This integration enables the management of complex systems in many domains, like energy distribution, smart buildings or manufacturing. While such systems facilitate a new way to face complex problems, their increasing complexity makes it difficult for users to understand the system's behavior. Therefore, the concept of ExpCPSs was developed to ensure understandability and trust even for increasingly complex systems [1].

ExpCPS research addresses critical challenges in CPS research, benefiting system users, engineers, and researchers alike. By offering user-centered and actionable explanations for system behaviors, the framework empowers users, such as facility operators, customers, and techni-

Proceedings of the Doctoral Consortium at ISWC 2024, co-located with the 23rd International Semantic Web Conference (ISWC 2024)

katrin.schreiberhuber@wu.ac.at (K. Schreiberhuber)

https://research.wu.ac.at/de/persons/katrin-schreiberhuber (K. Schreiberhuber)

0000-0003-1815-8167 (K. Schreiberhuber)

cians, to make informed decisions and enhance system performance. Engineers benefit from improved troubleshooting and optimization capabilities, while researchers gain a foundational infrastructure for further innovation. In essence, the ExpCPS framework advances transparency, understanding, and usability in CPS deployments, fostering safer and more efficient systems for all stakeholders.

As an example CPS, a smart grid intelligently manages energy consumption and production to ensure a stable grid operation. In times of high total consumption, the charging capabilities of an Electric Vehicle (EV) charging facility may be curtailed to avoid an energy overload at the local transformer. While this is a reasonable decision, an EV user only experiences longer charging times. Without an explanation, they would assume a system error and might refrain from using this charging facility in the future. In an ExpCPS, such an explanation can be generated by the system automatically, without the need for time-consuming data analysis by system experts. Such a capability does not only benefit the EV user in this scenario. It helps customer service representative to easily access comprehensive explanations for their customers. System managers can analyse common faults, identify error-prone devices and better design future smart grids in their long-term planning.

Some ExpCPS approaches have already been defined for some domains, such as smart grids [2, 3, 4, 5, 6] and smart buildings [7, 8]. As this research field is still relatively young, current solutions still face several research challenges: (1) Existing approaches are limited to domain-specific solutions. (2) The acquisition and integration of heterogeneous data sources poses a challenge as CPS data from different sources is relevant for explaining system behaviour (e.g., sensor measurements, topology data, domain knowledge) [9]. (3) Current explanation methods generate explanations tailored for AI/engineering experts, not end-users of a system [10].

2. Research Questions and Objectives

Based on previous work, our hypothesis is that SWT have the capabilities to address all of these challenges [11]. (1) Semantic modeling with ontologies facilitates explicit and formal representation of semantics in varying levels of detail, allowing for a generic, domain-independent representation of knowledge. (2) Its high degree of flexibility enables the unambiguous representation of semantics, enabling reasoning over heterogeneous data. (3) The representation of domain knowledge facilitates the conversion of technical explanations to match user vocabulary.

Therefore, the overarching research question we pose in this PhD is *"RQ0: How can SWT be used to enhance ExpCPS research?"*. This question translates to three core research questions, connected to more concrete research contributions, as shown in Fig. 1.

RQ1: What are knowledge representation needs for solving the ExpCPS problem with SWT?
Ontologies and ontology design patterns have already been developed to represent CPS data [12], events [13], causality [14] and explanations [15]. However, there is no unified model yet, which can be used out-of-the-box for the development of an ExpCPS.

C1: ExpCPS ontology. Our contribution is the development of a domain-independent ontology, which enables the integration of all data needed to explain system states of a CPS. It should provide a general data structure, which can be extended with domain-specific ontologies depending on the use case implementation.

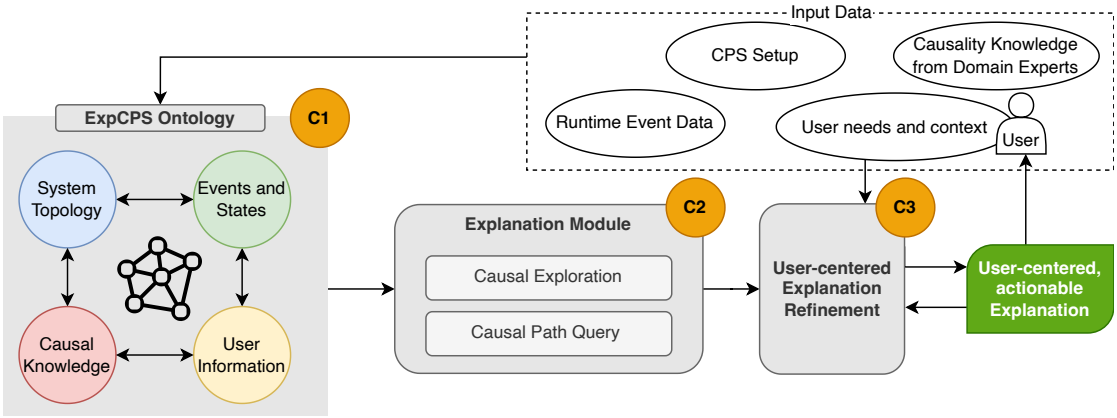

Figure 1: Research Contributions as components of an ExpCPS framework

RQ2: What are mechanisms to derive causal paths based on semantic data?
Previous CPS research has developed various methods for root cause analysis, such as Failure Mode and Effect Analysis (FMEA) [16] or Fault Tree Analysis (FTA) [17]. However, these methods are mostly used in the manufacturing domain, where fault analysis is conducted in fully engineered systems, meaning that most variables in a process can be measured and manipulated.

C2: Explanation Module. We aim to develop a causal reasoning module based on semantic data. We exploit expert knowledge about general causality relations, system setup knowledge as well as event data to derive actual causalities between system states. We will apply Knowledge Graph-based algorithms, extracting knowledge based on the ExpCPS ontology as defined in C1.

RQ3: What factors should be considered to incorporate user needs and context into CPS explanations? While the importance of user-centered explanations has been mentioned in various papers [18, 19, 20], there is limited research on how to achieve such user-centered explanations in CPSs.

C3: User-centered Explanation Refinement. We aim to develop a method to include user context in the creation of CPS explanations to ensure understandable and useful explanations for any system user. User needs and knowledge as well as user access rights are crucial aspects to construct a user-centered explanation. Each user has their individual goals as well as resources and rights, which influence the extent and version of information that is ideally presented to them.

3. Related Work

The research problem to be addressed in this PhD stems from the CPS research community, which includes various sub-communities, such as smart grid research, smart building research and industrial engineering. We aim to bring in solutions to these communities from the semantic

web research community, defining approaches to solve problems in novel application domains.

(Exp)CPS Research (Problem Space): In industrial CPS research, anomaly detection and subsequent root-cause analysis have been explored for more than 20 years. Common methods, such as FMEA [16] and FTA [17] are based on the specification of anomalies and corresponding causes by experts. These methods are heavily reliant on documents containing natural language descriptions, which potentially contain ambiguous and fragmented knowledge [21]. As systems are becoming increasingly complex and the number of system components is too large to be handled manually, thus a more automated approach is needed for future analysis. Additionally, the integration of various hetereogeneous data sources in an integrated representation still needs to be addressed (see RQ1).

While various approaches have addressed the representation of causal relations, such as attack trees [22] and fault trees [23]. These representations are domain-specific and struggle to incorporate domain knowledge to allow for automatic reasoning about causes of system states [24] (see RQ2).

The importance of user-centered explanations has been stated in various position papers [25, 18, 20]. In the European Commission's vision of Industry 5.0 (human-centric industrial systems that aim beyond efficiency and productivity), user-friendly and understandable explanations are identified as core features of such systems [26]. However, there are still limited approaches on how to solve this research problem (see RQ3).

Semantic Web Research (Solution Space): In Semantic Web Research, various resources have been developed to integrate heterogeneous data and knowledge. Ontology-based knowledge representations specifically for FMEA research have been proposed [27, 28] to address ambiguous data representation issues. For the representation of system topologies, the SOSA Ontology is a "lightweight general-purpose ontology to represent the interaction between entities" in CPSs [12]. As SOSA remains very general in the definition of systems and sensors, it can be extended with more domain-specific ontologies (e.g., Brick Ontology [29] for smart building domain) depending on the use case. To address the representation of events, causality and explanations, an ontology design pattern for representing causality was proposed in [14], as well as the Explanation Ontology [15] for user-centered AI explanations.

Initial research on the design of a human-centered ExpCPS has proposed a user-centered explanation engine in the smart home domain [30]. In their system, user profiles, roles and their current context are considered in the explanation generation process.

4. Research Methods and Evaluation

We aim to address our research questions using the design science methodology [31], along its three core cycles.

Relevance Cycle. We ensure the relevance of our artifacts by collecting user requirements from domain experts in a use case definition phase. As this PhD research project is funded by the Austrian national research agency (FFG), we make sure to align our research with the requirements from our use case partners at each step. Our use case partners come from the smart grid domain as well as the smart building domain. This ensures the coverage of requirements from different ExpCPS domains.

Research Cycle. Based on user requirements, we design concrete artifacts and processes to address the predefined problems (ExpCPS Ontology, Causal Reasoning Module, User-centered Explanation Refinement Module).

The evaluation of our research artifacts is conducted from multiple perspectives. On a technical level, an initial feasibility study is designed to evaluate the proposed ExpCPS components on their general ability to address the problem in different use cases. As a next step, a more rigorous evaluation will test the proposed system on its scalability and accuracy by applying it to a large-scale use case, checking the precision and recall of explained events based on ground truth knowledge (ground truth explanations are obtained by asking experts to give explanations as well as analysing time series data directly). Evaluation of user-centered explanations relies on more qualitative evaluation methods, involving system users to evaluate the correctness and understandability of explanations about the system they are using.

Concretely, we will test the proposed solutions in three different use cases from two domains. A small smart charging use case consists of an EV charging garage, which is connected to a local distribution grid. Anomalies of unexpected high energy consumption require explanations for a facility operator. This use case has a limited size, but contains real-life data, showing the feasibility of our approach in an in-vivo use case. A local energy community use case is defined in a simulation environment [32]. In this use case, in addition to the correctness of the solutions, the limits of the framework in terms of scalability are tested. Using a simulation environment for evaluation allows for an incremental increase in system sizes, requiring limited resources. In a third use case, our approaches are tested in a smart building environment, aiming to explain unusual heating behavior in an office building. Applying the approach in two different domains ensures the generalizability of the framework.

Rigor Cycle. All of our research is grounded in current state-of-the-art research, as already discussed in Section 3. For RQ1, the ontology engineering process was conducted using the LOT-methodology [33], ensuring a structured definition of competency questions. As much as possible, existing ontologies are extended to facilitate reuse of our methods in different domains and use cases. Existing causal models as well as causality representation methods are considered when addressing RQ2. For RQ3, current state of the art approaches are investigated on how to design user-centered explanations in multiple domains (e.g., XAI explanations, CPS explanations), as the research area is still very young.

5. Initial Results

In the first year of the PhD, we have already addressed RQ1 in a first iteration. Using the LOT methodology [33], we have collected user requirements and competency questions to design a first version of an ExpCPS ontology, called SENSE Ontology[1]. It is capable of integrating system topology data, event and state data as well as causal knowledge from domain experts. The ontology is currently an extension of the SOSA Ontology[2]. We propose an extension of the SOSA Ontology especially in adding causality knowledge between Observations in the represented system to allow causal reasoning over Events and States. In future iterations, this ontology

[1] http://w3id.org/explainability/sense#
[2] https://www.w3.org/TR/vocab-ssn/

will be extended to represent user-related information and explanation-related knowledge to facilitate user-centered explanations. The main evaluation is based on the applicability of the ontology for the two use cases of the research project (smart grid and smart building domain) as well as the ability of the ontology to answer the competency questions collected from system experts at the start of the ontology development process. Additionally, a structural evaluation of the ontology and its pitfalls can be conducted using OOPS [34].

Addressing RQ2, a first implementation of the Causal Reasoning Module was developed using SPARQL. In a two-step process, an algorithm first reasons over the system to explore actual causal knowledge in the data for knowledge base completion. In the second step, a query extracts the causal path from a root cause to a concrete trigger event, requiring an explanation.

In a feasibility study, the first implementation of the ExpCPS Ontology and the Explanation Module was tested on a small EV charging use case. The SENSE Ontology was able to represent all data needed for the explanation generation process, including system topology, events and states, as well as causal knowledge, which was defined a priori by domain experts. We are currently in the process of submitting a paper to the Energy Informatics community proposing the causal reasoning module as a solution to making smart grids more explainable.

We have not yet investigated the research area connected to RQ3 in detail. Thus, expanding the framework to consider user needs and context will be the next step to allow for a feasibility evaluation of the full framework. In addition, iterative improvements of each of the ExpCPS components shown in Fig. 1 will be conducted in future design science research cycles.

6. Reflection and Future Work

This research proposal shows the need for a domain-independent ExpCPS approach, which will be realized using SWT. Three research contributions are defined, which constitute components of a common ExpCPS framework. (i) An ExpCPS ontology will be designed to enable the integration of hetereogeneous data sources for creating explanations. (ii) An explanation module will exploit expert knowledge to derive actual causalities between system states. (iii) User-centered explanation refinement will include user context in the creation of explanations to ensure useful explanations, catered to a user's needs.

The goals of this PhD Project are closely aligned with the SENSE research project, which also enables the evaluation of the proposed solutions in a smart grid and a smart building use case.

As one limitation of this project, applications outside of these domains will not be covered during the PhD. This means, only a limited claim of applicability over all CPS domains can be formed. Additionally, the representation of causality was chosen to as an ontology. More specialized methods for causal representations and inference have been developed. While this design choice was consciously made to be more flexible in terms of knowledge integration, there could be restrictions on the capabilities of ontologies on the way. Furthermore, the integration of existing, automated explanation approaches has not been considered yet, and might be difficult to integrate.

In the next year, the goal is to have a first implementation of the full ExpCPS framework, including initial versions of all three contributions of this PhD. From this point on, a more rigorous evaluation can be conducted, analysing scalability and accuracy performance of the

framework. Depending on the evaluation results, improvements of all three components will be investigated in the third year of the PhD to create a robust, usable and accurate explainability framework for CPSs. Once the framework is envisioned, the application of sub-symbolic AI methods can be investigated to improve certain aspects of the system or assist experts in their knowledge discovery.

Acknowledgements

I would like to thank Dr. Marta Sabou and Dr. Fajar J. Ekaputra for their invaluable support and inputs during supervision. This work has been funded by the FFG SENSE project (project Nr. FO999894802).

References

[1] S. S. Jha, An Overview on the Explainability of Cyber-Physical Systems, The International FLAIRS Conference Proceedings 35 (2022). URL: https://journals.flvc.org/FLAIRS/article/view/130646. doi:10.32473/flairs.v35i.130646.

[2] P. R. Aryan, F. J. Ekaputra, M. Sabou, D. Hauer, R. Mosshammer, A. Einfalt, T. Miksa, A. Rauber, Simulation support for explainable cyber-physical energy systems, in: 2020 8th Workshop on Modeling and Simulation of Cyber-Physical Energy Systems, IEEE, 2020, pp. 1–6.

[3] P. R. Aryan, F. J. Ekaputra, M. Sabou, D. Hauer, R. Mosshammer, A. Einfalt, T. Miksa, A. Rauber, Explainable cyber-physical energy systems based on knowledge graph, in: Proceedings of the 9th Workshop on Modeling and Simulation of Cyber-Physical Energy Systems, 2021, pp. 1–6.

[4] A. Chhokra, N. Mahadevan, A. Dubey, G. Karsai, Qualitative fault modeling in safety critical cyber physical systems, in: Proceedings of the 12th System Analysis and Modelling Conference, 2020, pp. 128–137.

[5] J. Cordova, L. M. K. Sriram, A. Kocatepe, Y. Zhou, E. E. Ozguven, R. Arghandeh, Combined electricity and traffic short-term load forecasting using bundled causality engine, IEEE Transactions on Intelligent Transportation Systems 20 (2018) 3448–3458.

[6] J. E. Larsson, B. Öhman, A. Calzada, Real-time root cause analysis for power grids, in: Security and Reliability of Electric Power Systems, CIGRE Regional Meeting, Tallinn, Estonia, Citeseer, 2007.

[7] C. Lork, V. Choudhary, N. U. Hassan, W. Tushar, C. Yuen, B. K. K. Ng, X. Wang, X. Liu, An ontology-based framework for building energy management with iot, Electronics 8 (2019) 485.

[8] J. Ploennigs, A. Schumann, F. Lécué, Adapting semantic sensor networks for smart building diagnosis, in: The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part II 13, Springer, 2014, pp. 308–323.

[9] R. Minerva, G. M. Lee, N. Crespi, Digital twin in the iot context: A survey on technical features, scenarios, and architectural models, Proceedings of the IEEE 108 (2020) 1785–1824.

[10] R. Machlev, L. Heistrene, M. Perl, K. Levy, J. Belikov, S. Mannor, Y. Levron, Explainable artificial intelligence (xai) techniques for energy and power systems: Review, challenges and opportunities, Energy and AI 9 (2022) 100169.
[11] M. Sabou, S. Biffl, A. Einfalt, L. Krammer, W. Kastner, F. J. Ekaputra, Semantics for cyber-physical systems: A cross-domain perspective, Semantic Web 11 (2020) 115–124.
[12] K. Janowicz, A. Haller, S. J. Cox, D. Le Phuoc, M. Lefrançois, Sosa: A lightweight ontology for sensors, observations, samples, and actuators, Journal of Web Semantics 56 (2019) 1–10.
[13] Y. Raimond, S. Abdallah, The event ontology (2007).
[14] U. Jaimini, C. Henson, A. Sheth, An ontology design pattern for representing causality (2023).
[15] S. Chari, O. Seneviratne, M. Ghalwash, S. Shirai, D. M. Gruen, P. Meyer, P. Chakraborty, D. L. McGuinness, Explanation ontology: A general-purpose, semantic representation for supporting user-centered explanations, Semantic Web (2023) 1–31.
[16] M. Ben-Daya, Failure mode and effect analysis, in: Handbook of maintenance management and engineering, Springer, 2009, pp. 75–90.
[17] C. A. Ericson, Fault tree analysis primer, CreateSpace Incorporated, 2011.
[18] S. Dey, P. Chakraborty, B. C. Kwon, A. Dhurandhar, M. Ghalwash, F. J. S. Saiz, K. Ng, D. Sow, K. R. Varshney, P. Meyer, Human-centered explainability for life sciences, healthcare, and medical informatics, Patterns 3 (2022).
[19] F. Doshi-Velez, M. Kortz, R. Budish, C. Bavitz, S. Gershman, D. O'Brien, K. Scott, S. Schieber, J. Waldo, D. Weinberger, et al., Accountability of ai under the law: The role of explanation, arXiv preprint arXiv:1711.01134 (2017).
[20] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.
[21] L. Dittmann, T. Rademacher, S. Zelewski, Performing fmea using ontologies, in: 18th International Workshop on Qualitative Reasoning. Evanston USA, 2004, pp. 209–216.
[22] B. Schneier, Modeling security threats, Dr. Dobb's journal 24 (1999).
[23] B. Kordy, L. Piètre-Cambacédès, P. Schweitzer, Dag-based attack and defense modeling: Don't miss the forest for the attack trees, Computer science review 13 (2014) 1–38.
[24] A. Ibrahim, S. Kacianka, A. Pretschner, C. Hartsell, G. Karsai, Practical Causal Models for Cyber-Physical Systems, in: J. M. Badger, K. Y. Rozier (Eds.), NASA Formal Methods, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2019, pp. 211–227. doi:10.1007/978-3-030-20652-9_14.
[25] Roberto Confalonieri, Giancarlo Guizzardi, On the Multiple Roles of Ontologies in Explainable AI, Neurosymbolic Artificial Intelligence Journal (2023). URL: https://www.neurosymbolic-ai-journal.com/paper/multiple-roles-ontologies-explainable-ai.
[26] E. Commission, D.-G. for Research, Innovation, A. Renda, S. Schwaag Serger, D. Tataj, A. Morlet, D. Isaksson, F. Martins, M. Mir Roca, C. Hidalgo, A. Huang, S. Dixson-Declève, P. Balland, F. Bria, C. Charveriat, K. Dunlop, E. Giovannini, Industry 5.0, a transformative vision for Europe – Governing systemic transformations towards a sustainable industry, Publications Office of the European Union, 2021. doi:doi/10.2777/17322.
[27] Z. Rehman, C. V. Kifor, An ontology to support semantic management of fmea knowledge, International Journal of Computers Communications & Control 11 (2016) 507–521.

[28] A. Zhou, D. Yu, W. Zhang, A research on intelligent fault diagnosis of wind turbines based on ontology and fmeca, Advanced Engineering Informatics 29 (2015) 115–125.

[29] B. Balaji, A. Bhattacharya, G. Fierro, J. Gao, J. Gluck, D. Hong, A. Johansen, J. Koh, J. Ploennigs, Y. Agarwal, M. Berges, D. Culler, R. Gupta, M. B. Kjærgaard, M. Srivastava, K. Whitehouse, Brick: Towards a Unified Metadata Schema For Buildings, in: Proceedings of the 3rd ACM International Conference on Systems for Energy-Efficient Built Environments, ACM, Palo Alto CA USA, 2016, pp. 41–50. URL: https://dl.acm.org/doi/10.1145/2993422.2993577. doi:10.1145/2993422.2993577.

[30] M. Sadeghi, L. Herbold, M. Unterbusch, A. Vogelsang, Smartex: A framework for generating user-centric explanations in smart environments, arXiv preprint arXiv:2402.13024 (2024).

[31] A. R. Hevner, S. T. March, J. Park, S. Ram, Design science in information systems research, MIS quarterly (2004) 75–105.

[32] R. Mosshammer, K. Diwold, A. Einfalt, J. Schwarz, B. Zehrfeldt, Bifrost: A smart city planning and simulation tool, in: Intelligent Human Systems Integration 2019: Proceedings of the 2nd International Conference on Intelligent Human Systems Integration (IHSI 2019): Integrating People and Intelligent Systems, February 7-10, 2019, San Diego, California, USA, Springer, 2019, pp. 217–222.

[33] M. Poveda-Villalón, A. Fernández-Izquierdo, M. Fernández-López, R. García-Castro, Lot: An industrial oriented ontology engineering framework, Engineering Applications of Artificial Intelligence 111 (2022) 104755.

[34] M. Poveda-Villalón, A. Gómez-Pérez, M. C. Suárez-Figueroa, Oops!(ontology pitfall scanner!): An on-line tool for ontology evaluation, International Journal on Semantic Web and Information Systems (IJSWIS) 10 (2014) 7–34.