An Ontology-Enabled Approach For User-Centered and Knowledge-Enabled Explanations of AI Systems

Shruthi Chari[1] [0000−0003−2946−7870]

Rensselaer Polytechnic Institute, Troy NY 12180, USA

Abstract. Explainable Artificial Intelligence (AI) focuses on helping humans understand the working of AI systems or their decisions and has been a cornerstone of AI for decades. Recent research in explainability has focused on explaining the workings of AI models or model explainability. There have also been several position statements and review papers detailing the needs of end-users for user-centered explainability but fewer implementations. Hence, this thesis seeks to bridge some gaps between model and user-centered explainability. We create an explanation ontology (EO) to represent literature-derived explanation types via their supporting components. We implement a knowledge-augmented question-answering (QA) pipeline to support contextual explanations in a clinical setting. Finally, we are implementing a system to combine explanations from different AI methods and data modalities. Within the EO, we can represent fifteen different explanation types, and we have tested these representations in six exemplar use cases. We find that knowledge augmentations improve the performance of base large language models in the contextualized QA, and the performance is variable across disease groups. In the same setting, clinicians also indicated that they prefer to see actionability as one of the main foci in explanations. In our explanations combination method, we plan to use similarity metrics to determine the similarity of explanations in a chronic disease detection setting. Overall, through this thesis, we design methods that can support knowledge-enabled explanations across different use cases, accounting for the methods in today's AI era that can generate the supporting components of these explanations and domain knowledge sources that can enhance them.

Keywords: Explainable AI · Knowledge-Enabled Explanations · User-Centered Explanations.

1 Introduction and Problem statement

Artificial Intelligence (AI) has evolved over the years from having limited applications in application domains such as the military to having more widespread use to being available to assist humans in both high-precision tasks such as healthcare and financial decisions to more everyday tasks such as helping in web search, weather alerts to navigation. Through these applications and improvements in computing and AI technology, AI methods have evolved to support the various applications and better computing infrastructure. Also, through these AI evolutions, different approaches have emerged, from rule-based expert systems to pattern-based machine learning (ML) and deep learning methods.

As humans, we tend to trust AI better if we can understand the reasoning of how the AI came to a decision or connect an AI decision to what we are familiar with [21] [15]. Understandably so, explainability or explainable AI (XAI) has been one of the earliest conceived thrusts of what we know today as Trustworthy AI [20], from early works such as Mycin [19] that were developed to explain an expert system in a medical setting to today the plethora of post-hoc explainability methods that provide reasoning for features that were found to be crucial by somewhat opaque ML models. The needs for explainability are diverse and are evolving in the changing AI landscape [8] [9]. For example, as humans, we reason through different paths before we trust a decision, i.e., explanations can be multi-dimensional. Additionally, explanations are typically reactive to user questions [7] [10] [8], and often have multiple forms and types such as the what ifs or counterfactuals, what evidence or scientific and what data or data-based [9] [16].

In recent times, given the increase in complexity of AI and ML methods; much of the explainability research has focused on model explanations [2] [17] alone. However, several researchers [15] [16] [9] have posited the need for conversational user-centered explainability beyond model explanations alone, which is of multiple types and supported by various sources such as data, knowledge, and context. Many publications in user-centered explainability have either been position statements [7] [8] [15] or survey papers [24], with fewer implementations that can be applied across use cases [13]. Hence, there are opportunities within explainable AI to support user-centered explanations that build from different data sources and AI methods and can address a range of user questions. Upon laying out the research questions that can tackle some of these gaps (Sec. 2), I will describe the contributions (Sec. 2.2) we have made; focusing on methods we use or plan to use to support these contributions and then describing the results (Sec. 3) thus far from them.

2 Research Questions and Contributions

2.1 Research Questions

This thesis hopes to address the following research questions in the evolving field of user-centered explainable AI:

- How can we formally represent explanations with support for interacting AI systems, additional data sources, and along different dimensions?
- How useful and feasible are such explanations for clinical settings?
- Is it feasible to combine explanations from multiple data modalities and AI methods?

2.2 Contributions and Methods

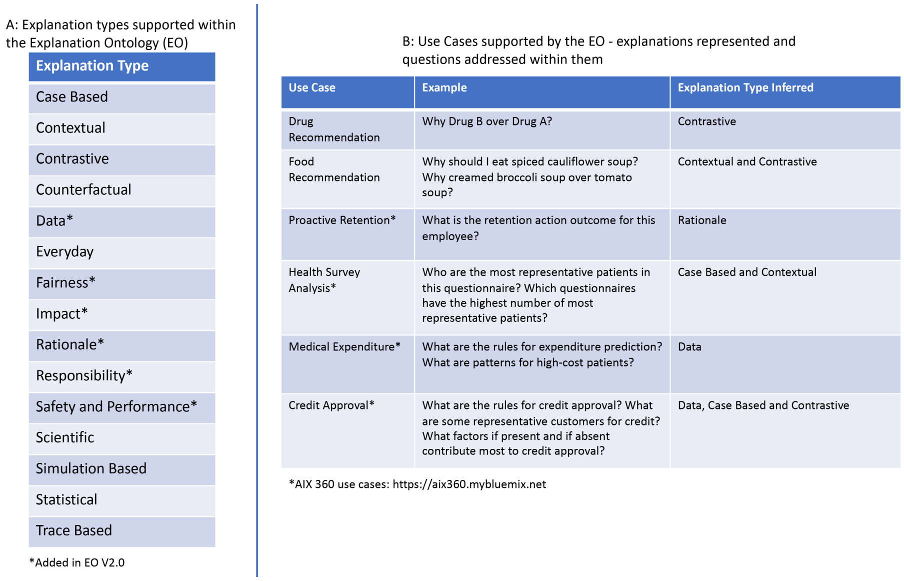
Here are the contributions that address the previously listed research questions.

- Explanation Ontology: We design an Explanation Ontology (EO), a general-purpose semantic representation that can represent fifteen different literature-derived explanation types via their system-, interface- and user- related components [6]. We showcase the utility of the EO's model to represent explanation types across six different use cases ranging from domains of finance, food to healthcare. We design competency questions that our target end-user, a system developer, would ask when using the EO. Further, we have released two versions of the EO with, with added support for an evolving hybrid AI landscape by by including a wider range of commonly used explainer methods in EO V2.0. The EO is open-sourced and available at: https://tetherless-world.github.io/explanation-ontology/index.
- Contextualizing Model Explanations via a Knowledge Augmented Question-Answering Method: We design a clinical question-answering (QA) system to address questions from clinical practice guidelines (CPGs) to provide contextual explanations to help clinicians interpret risk prediction scores and their post-hoc explanations in a comorbidity risk prediction setting [5]. We refined the use case in consultation with a clinician. We identified dimensions of interest in the use case, along with which contextual explanations would be helpful for clinicians to interpret the risk scores and patient features better - patient, their predicted risk and post-hoc explanations of their risk. From an implementation standpoint, we developed a QA framework to extract and support contextual explanations from CPGs. We leverage large language models (LLMs) and their clinical variants for the QA and build knowledge augmentations (KAs) to improve the semantic coherence of the answers to the questions. We conduct a quantitative evaluation to demonstrate the QA's efficacy and a qualitative evaluation with clinicians to understand if contextual explanations are helpful and where else they would require support to use them in their practice.
- A Method to Combine Multiple Explanations: We are designing a general-purpose framework capable of providing multiple explanations to an end-user question (e.g., that of a clinician). Within the framework, we want to break down a user question into sub-questions that can be addressed by different explanation types either those supported by different explainer methods or different data modalities. As a second step, we are developing a method to combine explanations if they supplement each other and leave them as is if not. Overall, we hope to generate natural language explanations from their individual data, knowledge, and method output components. We plan to evaluate the explanations via small-scale user studies and use / suggest metrics [27] that the XAI community has proposed or will benefit from.

3 Results and Evaluations

3.1 Explanation Ontology

We demonstrate how the EO can represent explanations via their dependencies in six different use cases (Fig. 1), where explanations are generated by either IBM's AIX-360 suite of explainer methods [2] or via other self-explainable logical reasoners (i.e., food and drug recommendation use cases). Upon running a reasoner on these use case knowledge graphs (KGs), we can also infer different explanation types supported within the EO, making it possible for system designers to generate various explanation types within their use cases. Furthermore, we borrow ontology evaluation techniques from Muhammad et al. [1] and adopt three different evaluation strategies from their paper, i.e., evolution-based evaluation to highlight the benefits of the additions made in EO V2.0, task-based evaluation to demonstrate what general support system designers can expect to seek when exploring EO or planning to use EO in their use cases and application-based evaluation to indicate what kinds of questions can be asked around EO supported use case KGs. Our published papers on the EO [6] can provide more details on our assessment and results.



A: Explanation types supported within the Explanation Ontology (EO)

Explanation Type
Case Based
Contextual
Contrastive
Counterfactual
Data*
Everyday
Fairness*
Impact*
Rationale*
Responsibility*
Safety and Performance*
Scientific
Simulation Based
Statistical
Trace Based

*Added in EO V2.0

B: Use Cases supported by the EO - explanations represented and questions addressed within them

Use Case	Example	Explanation Type Inferred
Drug Recommendation	Why Drug B over Drug A?	Contrastive
Food Recommendation	Why should I eat spiced cauliflower soup? Why creamed broccoli soup over tomato soup?	Contextual and Contrastive
Proactive Retention*	What is the retention action outcome for this employee?	Rationale
Health Survey Analysis*	Who are the most representative patients in this questionnaire? Which questionnaires have the highest number of most representative patients?	Case Based and Contextual
Medical Expenditure*	What are the rules for expenditure prediction? What are patterns for high-cost patients?	Data
Credit Approval*	What are the rules for credit approval? What are some representative customers for credit? What factors if present and if absent contribute most to credit approval?	Data, Case Based and Contrastive

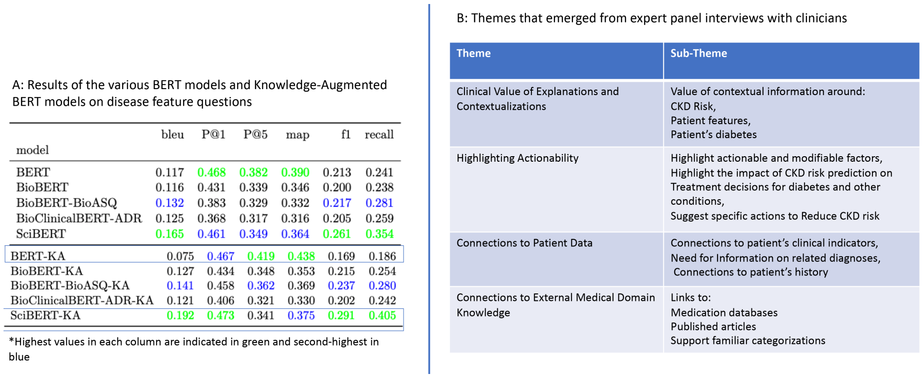
*AIX 360 use cases: https://aix360.mybluemix.net

Fig. 1. Explanation types supported within and Use Cases represented by the EO.

3.2 A Method to Support Contextual Explanations

We implement our QA pipeline to support contextual explanations in the risk prediction of a comorbidity of type-2 diabetes - chronic kidney disease (CKD).

We extract content from the then current edition of the type-2 diabetes CPG or American Diabetes Association (ADA) 2021 CPG [4]. We provide contextual explanations for five different question types including questions about the patient's diabetes and risk summaries, questions about important features contributing to their risk (typically found to be other diseases they have) and their medication and lab values. Each of these question types provide context for different entities of interest, e.g., the patient's risk summary contextualizes their risk for CKD. Our KA QA pipeline provides answers to three of these question types from the ADA CPG - feature importance questions, medications, and lab value questions. For the KA settings of the QA, we leverage knowledge from medical coding schemes - Snomed-CT and UMLS. We evaluate the answers from the KA QA models via standard QA evaluation metrics (F1, mean average precision (MAP) and recall) and find improvements in precision using KAs to the LLMs (Fig. 2). We also perform additional analyses to understand if any LLM architecture or KA strategy performs much better at certain disease groups than others and we find that there are differences in performance within disease groups. Additionally, we categorize our conversations during the expert panel sessions with clinicians into themes and sub-themes (Fig. 2); where they indicated that they preferred support from explanations or would want to see more support. During these sessions, we walked clinicians through contextual explanations for prototypical patients on a risk prediction dashboard that we developed. More details are available in our paper [5].



A: Results of the various BERT models and Knowledge-Augmented BERT models on disease feature questions

model	bleu	P@1	P@5	map	f1	recall
BERT	0.117	0.468	0.382	0.390	0.213	0.241
BioBERT	0.116	0.431	0.339	0.346	0.200	0.238
BioBERT-BioASQ	0.132	0.383	0.329	0.332	0.217	0.281
BioClinicalBERT-ADR	0.125	0.368	0.317	0.316	0.205	0.259
SciBERT	0.165	0.461	0.349	0.364	0.261	0.354
BERT-KA	0.075	0.467	0.419	0.438	0.169	0.186
BioBERT-KA	0.127	0.434	0.348	0.353	0.215	0.254
BioBERT-BioASQ-KA	0.141	0.458	0.362	0.369	0.237	0.280
BioClinicalBERT-ADR-KA	0.121	0.406	0.321	0.330	0.202	0.242
SciBERT-KA	0.192	0.473	0.341	0.375	0.291	0.405

*Highest values in each column are indicated in green and second-highest in blue

B: Themes that emerged from expert panel interviews with clinicians

Theme	Sub-Theme
Clinical Value of Explanations and Contextualizations	Value of contextual information around: CKD Risk, Patient features, Patient's diabetes
Highlighting Actionability	Highlight actionable and modifiable factors, Highlight the impact of CKD risk prediction on Treatment decisions for diabetes and other conditions, Suggest specific actions to Reduce CKD risk
Connections to Patient Data	Connections to patient's clinical indicators, Need for Information on related diagnoses, Connections to patient's history
Connections to External Medical Domain Knowledge	Links to: Medication databases Published articles Support familiar categorizations

Fig. 2. Quantitative results on disease questions and qualitative results from our contextual explanations

3.3 A Method to Combine Explanations

We are currently implementing methods to generate and analyze explanations from different data modalities. We are evaluating this implementation in a multi-modality disease setting, such as the staging of CKD. We are generating explanations for the patients' genetics and their CT scans and we want to answer if these explanations from two different data modalities complement / supplement each other. We are using text similarity metrics as a start to evaluate the degree of agreement of the explanations and might consider using other XAI metrics. We are working towards completing this contribution within a year. Also, we are submitting a patent disclosure on this work.

4 Related Work

We review related works in three different areas of contributions of this thesis including representations of model explanations, support for contextual explanations in clinical settings, and methods to combine explanations from various sources and methods. There are two ontologies to represent explanations [23] [22], one [23] focuses on the dependencies of explanations from the social sciences and the philosophy domains (we include classes from their ontology in the EO) and another [22] is a representation to support various explanation types in a specific engineering architecture, but this ontology is not in the standard OWL format. There have been papers [25] [3] on the need to support contextual explanations in an end-end setting where explanations provide context around entities of interest in an implemented setting such as risk prediction. Still, the implementations are few [18] [26]. Furthermore, the applications of LLMs to CPGs have been limited to non-QA tasks such as natural language understanding and entity recognition [11]. Architectures to combine explanations from different data sources and logical and statistical reasoners have been proposed and implemented in the past [14], but not in today's hybrid AI landscape. However, a few recent papers propose or design methods to identify if feature-based explanations agree or disagree with each other [12] and suggest building towards multi-input, conversational explanations. Attempts to combine or support various user-centered and natural language explanation types need to be improved, and hence there are opportunities and need for the explanation framework.

5 Conclusions

We have described three contributions that together enable the generation of explanations from various supporting components. The implementations for the first contribution, explanation ontology, and the last planned contribution are kept general-purpose and use case agnostic. However, we focus the second contribution on providing contextual explanations that offer additional domain knowledge to interpret model explanations and AI method results, in a clinical use case. Still, our methods can be adapted to other literature-rich settings. In each of our contributions, we leverage ontologies by building them ourselves or repurposing and utilizing well-used domain ontologies and knowledge graphs (KGs) in the field of use. This combination of neural and symbolic approaches within our contributions to support explanations helps root the explanations in domain knowledge, i.e., knowledge-enabled explanations and also makes the explanations more easily interpretable by domain experts. We have presented themes that clinicians indicated in their evaluation of our contextual explanations, and these can help inform future goals that explanations should focus on. Further, we have open-sourced the EO and have iterated through the ontology to support more explanation types and use cases, making the resource more accessible to our intended user group of system developers. Some of the themes from our discussions with clinicians and our experience including ontologies and KGs in explanations, can be valuable for discussion with other semantic web researchers. Overall, this thesis strides towards allowing different explanation types to be supported by a broad range of AI methods and knowledge sources while accounting for user requirements, attempting to make explanations more user-centered and multi-perspective.

Acknowledgements

This work is done as part of the HEALS project, and is partially supported by IBM Research AI through the AI Horizons Network. I thank my advisor, Prof. Deborah L. McGuinness and co-advisor, Prof. Oshani Seneviratne, whose inputs and guidance greatly helped assist and shape this thesis.

References

1. Amith, M., He, Z., Bian, J., Lossio-Ventura, J.A., Tao, C.: Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. Journal of biomedical informatics **80**, 1–13 (2018)
2. Arya, V., Bellamy, R.K., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S.C., Houde, S., Liao, Q.V., Luss, R., Mojsilović, A., et al.: Ai explainability 360: Impact and design. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 12651–12657 (2022)
3. Avgerou, C.: Contextual explanation: Alternative approaches and persistent challenges. MIS Quarterly **43**(3), 977–1006 (2019)
4. Care, D.: Standards of medical care in diabetes 2021. Diabetes Care **44**(Suppl 1) (2021)
5. Chari, S., Acharya, P., Gruen, D.M., Zhang, O., Eyigoz, E.K., Ghalwash, M., Seneviratne, O., Saiz, F.S., Meyer, P., Chakraborty, P., et al.: Informing clinical assessment by contextualizing post-hoc explanations of risk prediction models in type-2 diabetes. Artificial Intelligence in Medicine **137**, 102498 (2023)
6. Chari, S., Seneviratne, O., Ghalwash, M., Shirai, S., Gruen, D.M., Meyer, P., Chakraborty, P., McGuinness, D.L.: Explanation ontology: A general-purpose, semantic representation for supporting user-centered explanations. Semantic Web Journal p. In Press (2023)
7. Dey, S., Chakraborty, P., Kwon, B.C., Dhurandhar, A., Ghalwash, M., Saiz, F.J.S., Ng, K., Sow, D., Varshney, K.R., Meyer, P.: Human-centered explainability for life sciences, healthcare, and medical informatics. Patterns **3**(5), 100493 (2022)
8. Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Schieber, S., Waldo, J., Weinberger, D., Wood, A.: Accountability of ai under the law: The role of explanation. arXiv preprint arXiv:1711.01134 (2017)

9. Gilpin, L.H., Bau, D., Yuan, B.Z., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An approach to evaluating interpretability of machine learning. arXiv preprint arXiv:1806.00069 (2018)
10. Gunning, D.: Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web **2** (2017)
11. Hussain, M., Hussain, J., Ali, T., Ali, S.I., Bilal, H.S.M., Lee, S., Chung, T.: Text classification in clinical practice guidelines using machine-learning assisted pattern-based approach. Applied Sciences **11**(8), 3296 (2021)
12. Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., Lakkaraju, H.: The disagreement problem in explainable machine learning: A practitioner's perspective. arXiv preprint arXiv:2202.01602 (2022)
13. Lakkaraju, H., Slack, D., Chen, Y., Tan, C., Singh, S.: Rethinking explainability as a dialogue: A practitioner's perspective. arXiv preprint arXiv:2202.01875 (2022)
14. McGuinness, D.L., Glass, A., Wolverton, M., Da Silva, P.P.: Explaining task processing in cognitive assistants that learn. In: AAAI Spring Symp.: Interaction Challenges for Intelligent Assistants. pp. 80–87 (2007)
15. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
16. Mittelstadt, B., Russell, C., Wachter, S.: Explaining explanations in ai. In: Proc. of the Conf. on fairness, accountability, and transparency. pp. 279–288. ACM (2019)
17. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int. Conf. on knowledge discovery and data mining. pp. 1135–1144. ACM (2016)
18. Rieger, L., Singh, C., Murdoch, W., Yu, B.: Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In: International Conference on Machine Learning. pp. 8116–8126. PMLR (2020)
19. Shortliffe, E.H.: Mycin: a rule-based computer program for advising physicians regarding antimicrobial therapy selection. Tech. rep., STANFORD UNIV CALIF DEPT OF COMPUTER SCIENCE (1974)
20. of Standards, N.I., Technology: Trustworthy AI: Managing the Risks of Artificial Intelligence (2022), https://www.nist.gov/speech-testimony/trustworthy-ai-managing-risks-artificial-intelligence
21. Swartout, W., Paris, C., Moore, J.: Explanations in knowledge systems: Design for explainable expert systems. IEEE Expert **6**(3), 58–64 (1991)
22. Teze, J.C.L., Paredes, J.N., Martinez, M.V., Simari, G.I.: Engineering user-centered explanations to query answers in ontology-driven socio-technical systems. Semantic Web Journal: Ontolofies in XAI **In review**
23. Tiddi, I., d'Aquin, M., Motta, E.: An ontology design pattern to define explanations. In: Proceedings of the 8th Int. Conf. on Knowledge Capture. pp. 1–8 (2015)
24. Tiddi, I., Schlobach, S.: Knowledge graphs as tools for explainable machine learning: A survey. Artificial Intelligence **302**, 103627 (2022)
25. Tonekaboni, S., Joshi, S., McCradden, M.D., Goldenberg, A.: What clinicians want: contextualizing explainable machine learning for clinical end use. In: Machine learning for healthcare conference. pp. 359–380. PMLR (2019)
26. Zhang, X., Qian, B., Li, Y., Cao, S., Davidson, I.: Context-aware and time-aware attention-based model for disease risk prediction with interpretability. IEEE Transactions on Knowledge and Data Engineering (2021)
27. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5), 593 (2021)