Shruthi Chari[1] (charis@rpi.edu), Oshani Seneviratne[1] (senevo@rpi.edu), Daniel M. Gruen[2] (dangruen@gmail.com), Morgan A. Foreman[2] (morgan_foreman@ibm.com), Amar K. Das[2] (amardas@us.ibm.com), Deborah L. McGuinness[1] (dlm@cs.rpi.edu)

[1]Rensselaer Polytechnic Institute, Troy, NY

[2]IBM Research, Cambridge, MA

Supporting User-Centric Explanation Types for Clinical Reasoning

Classifications

Types of Computational Biomedical Knowledge; Systems, Platforms, Tools and Services

Background

Explainable Artificial Intelligence (AI) is receiving attention due to the increased proliferation of machine learning methods in high-precision settings. Traditionally, different methods in AI have tackled explainability from different angles tightly coupled with their capabilities. However, with the increasing adoption of AI, there is a need for **user-centric focus to explainability** that is urging researchers to think beyond the **"one explanation fits all"** [1] paradigm and explore methods for **"enhanced explainability"** [2] that consider what needs to be explained and in what setting.

Motivations

- Since, explanations need to adapt to **users' needs and contexts**, and various situations, we began reviewing literature for different explanation types.
- We found a lack of **infrastructure and support** to generate user-centric explanations that address a broad range of user questions (e.g., Why, Why Not, What Ifs, What Other, etc.)
- Further, there is a lack of **consensus on the definitions and components** of explanations and explanation types, which points to the need for a semantic representation.

Methods

- We found **nine distinct explanation types** in the literature [3] that have different strengths, rationales and serve different purposes. We redefined these explanation types along with a prototypical question that can be addressed by them (Table 1)
- We conducted a **user-centered design study with clinicians** to understand the usage of these explanation types in their practice.
- We designed an **Explanation Ontology** to model the role of explanations, both from a system and user attribute process, and the range of literature-derived explanation types (Fig. 1).

Discussion

In our approach, we have:

- Been able to utilize our Explanation Ontology to encode the **generational needs of explanation types**, gathered from an analysis of various components necessary to assemble these types from the literature as well as from our user study
- Have designed a **selected set of competency questions** to guide system developers about the intended use of our Explanation Ontology
- Found that some **explanation types are used more often** than others depending on the use case:
  - During our user study, clinicians were most often using **contextual explanations** and their experiential knowledge, **clinical pearls**, a form of everyday explanations

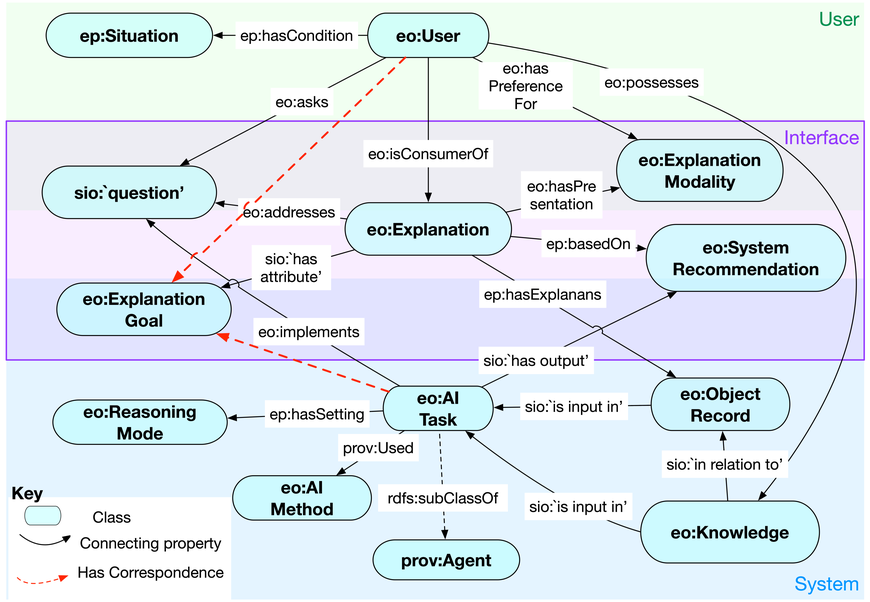

Fig. 1: A conceptual overview of our **Explanation Ontology**, capturing entities to allow explanations to be assembled by an **AI Task**, used in a system interacting with a user. We depict user-attributes of explanations in the upper portion (green highlight), system-attributes in the lower portion (blue highlight), and attributes that would be visible in a user interface are depicted in the middle portion in purple.

References

1. B. Mittelstadt, C. Russell, and S.Wachter, "Explaining explanations in AI," in Proc. of the Conf. on Fairness, Accountability, and Transparency. ACM, 2019, pp. 279–288
2. Matheny, M., Israni, S. T., Ahmed, M., & Whicher, D. (2020). Artificial intelligence in health care: The hope, the hype, the promise, the peril. *Natl Acad Med*, 94-97.
3. S. Chari, D. Gruen, O. Seneviratne, D. L. McGuinness, "Directions for Explainable Knowledge-Enabled Systems". In: Ilaria Tiddi, Freddy Lecue, Pascal Hitzler (eds.), Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges. Studies on the Semantic Web, IOS Press, Amsterdam, 2020, 245 - 261

Table 1: Catalog of literature-derived **Explanation Types**, where we present a clinically-oriented question that can be addressed by an explanation type

Explanation Type	Definition
Case-based	**What other situations with complex patients have had this recommendation applied?**
Contextual	**What broader information about the current situation prompted you to suggest this recommendation now?**
Contrastive	**Why administer this new drug over the one I would typically prescribe?**
Counterfactual	**What if the patient had a high risk for cardiovascular disease? Would you still recommend the same treatment plan?**
Everyday	**What are the signs I should be careful to check for in this case?** More specific form we identified in user study: *clinical pearls*
Scientific	**What is the biological basis, particularly the evidence, for this recommendation?** Further subclasses we identified in user study: *evidence-based* and *mechanistic*
Simulation-based	**What would happen if we prescribe this drug to the patient?**
Statistical	**What percentage of similar patients who received this treatment recovered?**
Trace-based	**What steps were taken (rules were fired) by the system to generate this recommendation?**

View more at: https://tetherless-world.github.io/explanation-ontology/

Take-away:

Our ontology-enabled approach can help:

- AI system designers to **design hybrid AI models that support different forms of reasoning that can generate different explanation types** which address user's needs gathered from user studies

Acknowledgments

This work is partially supported by IBM Research AI through the AI Horizons Network. We thank our colleagues from RPI, Sabbir Rashid, and, IBM Research, Ching-Hua Chen, who greatly assisted the research.