Shruthi Chari[1] (charis@rpi.edu), Oshani Seneviratne[1] (senevo@rpi.edu), Daniel M. Gruen[2] (daniel_gruen@us.ibm.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

Explanation Ontology: A Model of Explanations for User-Centered AI

Keywords
Explainable AI; Modeling of Explanations and Explanation Types; Supporting Explainable AI in Clinical Decision Making

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 explore fields of explanation sciences [1, 2] such as social sciences, philosophy, and computer science to avoid the "**one explanation fits all**" issue.

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 [2] 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

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

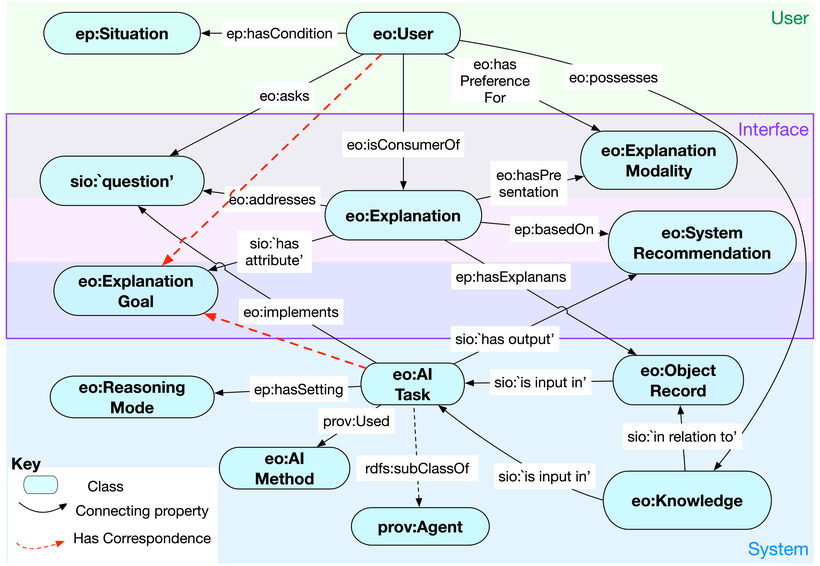

Fig. 1: A conceptual overview of our **Explanation Ontology**, capturing entities to allow explanations to be assembled by an **AI Task**, used in a system interacting with a user. We depict user-attributes of explanations in the upper portion (green highlight), system-attributes in the lower portion (blue highlight), and attributes that would be visible in a user interface are depicted in the middle portion in purple. Ontology prefixes used in this diagram expand to **eo**: Explanation Ontology, **ep**: Explanations Pattern Ontology, **prov**: The Provenance Ontology and **sio**: SemanticScience Integrated Ontology.

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

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

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

Explanation Type	Definition
Case-based	**What other situations with complex patients have had this recommendation applied?** Case-based explanations contain results that are based on actual prior cases that can be presented to the user to provide compelling support for the system's conclusions. These types of explanations can involve analogical reasoning, relying on similarities between features of the case and of the current situation.
Contextual	**What broader information about the current situation prompted you to suggest this recommendation now?** Contextual explanations are those that refer to information about items other than the explicit inputs and output, such as information about the user, situation, and broader environment that affected the computation.
Contrastive	**Why administer this new drug over the one I would typically prescribe?** Contrastive explanations define an output of interest and present contrasts between the fact (the event that did occur), the given output, and the foil (the event that did not occur), the output of interest.
Counterfactual	**What if the patient had a high risk for cardiovascular disease? Would you still recommend the same treatment plan?** Counterfactual explanations address the question of what results would have been obtained with a different set of inputs than those used.
Everyday	**What are the signs I should be careful to check for in this case?** Everyday explanations are accounts of the real world that appeal to the user based on their general understanding and knowledge.
Scientific	**What is the biological basis, particularly the evidence, for this recommendation?** Scientific explanations reference the results of rigorous scientific methods, such as observations and measurements.
Simulation-based	**What would happen if we prescribe this drug to the patient?** Simulation-based explanations are those that use an imagined or implemented imitation of a system or process and the results that emerge from similar inputs.
Statistical	**What percentage of similar patients who received this treatment recovered?** Statistical explanations present an account of the outcome based on data about the occurrence of events under specified (e.g., experimental) conditions. Statistical explanations refer to numerical evidence on the likelihood of factors or processes influencing the result.
Trace-based	**What steps were taken (rules were fired) by the system to generate this recommendation?** Trace-based explanations describe the underlying sequence of steps used by the system to arrive at a specific result. These types of explanations contain the line of reasoning per case and addresses the question of why and how the application did something.

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.