Ontology Summit 2019 Communiqué: Explanations

1 Introduction

The last few years have seen a substantial increase in the use of machine learning (ML) techniques for solving important problems in the field of Artificial Intelligence (AI). These advances have been driven by the availability of massive amounts of raw data. The ML techniques construct complex statistical models by processing large datasets. Unfortunately, it is difficult to explain how these ML models come to their conclusions, since each decision is, in principle, the result of a program that includes the entire dataset that was used to develop the model. This has led to the perception that ML models are too "deep" and "mysterious" to be adequately explained. Whether or not this is an accurate perception, to solve the problem of explaining the functioning of an ML model or, indeed, any large complex system, it is necessary to address the issue of what an explanation is, as well as what criteria can be used to evaluate the accuracy of an explanation and its suitability for a purpose.

In a thought-provoking presentation, Derek Doran asked the question "Okay but Really... What is Explainable AI?" He found that there was wide disagreement about the answer to this question. As a first approximation to classifying this variety, he identified a hierarchy of explanation "types" an AI system can exhibit:

1. Interpretable: I can identify why an input goes to an output.
2. Explainable: I can explain how inputs are mapped to outputs.
3. Reasonable: I can conceptualize how inputs are mapped to outputs.

While the first two types above may be adequate in some limited technical contexts, for the most part, humans would demand the third and highest level. Accordingly, there was general agreement with the following definition of explanation, "An explanation is the ability to answer a how or why question and to answer a follow-up question to clarify in a particular context." This definition makes it clear that an explanation can never be enough if it consists of a single generated answer, no matter how elaborate it might be. There must be a capability for a dialog between the user and the system. A further consequence of this definition is the importance of context. Since the notion of context was the subject of the Ontology Summit 2018, the current summit may be regarded as a continuation of the previous summit. Indeed, since context includes the answers to all of the "six Ws" (namely, Who, What, When, Where, Why and How), explanation is both part of the context and depends on it.

The Ontology Summit 2019 sought to explore, identify and articulate how ontology can bring value to the problem of automating explanations of complex systems in general. The summit dealt with this goal

by first studying the notion of explanation in a series of sessions in the Fall of 2018. The two most important areas that were identified at this time were (1) Commonsense Reasoning and Knowledge and (2) Narratives. These were then studied at greater depth in subsequent sessions in 2019. In addition, it was decided that some specific domains should be studied to determine the kinds of problems that practitioners are facing with respect to explanations. The two chosen domains were the Financial and the Medical domain, both of which were broadly defined. Finally, since explanations for AI were the original motivation, the problem of Explainable AI (XAI) was also studied. Accordingly, the summit consisted of sessions devoted to five tracks.

The purpose of this Communiqué is to identify some of the prevailing viewpoints and the major issues and challenges of explanation, especially the role that ontology could play toward solving these challenges. The Communiqué begins with some background and history of the notion of explanation in the next session. This is followed by sections devoted to synthesizing the findings from each of the five tracks. The Communiqué then presents findings, recommendations and conclusions.

2 Background

An explanation is the answer to the question "Why?" as well the answers to follow-up questions such as "Where do I go from here?" Accordingly, explanations generally occur within the context of a process, which could be a dialog between a person and a system or could be an agent-to-agent communication process between two systems. Explanations also occur is social interactions when clarifying a point, expounding a view, or interpreting behavior. In all such circumstances in common parlance one is giving/offering an explanation.

A brief history of *why* explanations provides some context and includes the observation that among the first known attempts at understanding the why of explanations were those documented among Greek and Indian intellectuals and philosophers. For example, to understand and explain the why there was a Peloponnesian War Thucydides defined explanations as a process where facts (indisputable data), which are observed, evaluated based on some common knowledge of human nature. This was then compared in order to reach generalized principles for *why* some events occur. In the writings of Plato (e.g., Phaedus and Theaetetus), we see explanations as an expression using logos knowledge composable by Universal Forms, which are abstractions of the world's entities we come to experience and know. Facts, in this view are occurrences or states of affairs and may be a descriptive part of an explanation, but not the deep Why. Aristotle's view, such as in Posterior Analytics provides a more familiar view of explanation as part of a logical, deductive, process using reason to reach conclusions. Aristotle proposed 4 types of causes (Aitia) to explain things. These were from either the thing's matter, form, end, or change-initiator (efficient cause). Following Descartes, Leibniz and especially Newton, modern deterministic causality using natural mechanisms became central to causal explanations. To know what causes an event means to employ natural laws as the central means to understand and explain why it happened. As this makes clear some notions of the nature of knowledge, how we come to know and the nature of reality are part of explanation.

The Ontology Summit 2019 was concerned with the role of ontologies for explaining the reasoning of a system. More specifically, the summit focused on critical explanation gaps and the role of ontologies for dealing with these gaps. The sessions examined current technologies and real needs driven by risks and requirements to meet legal or other standards.

Inspired by the current DARPA Explainable AI (XAI) Project, (see: [5, 6]) the Ontology Summit theme considered the general problem of explanation. The summit considered not only AI systems that can explain their actions, but also other smart engineering systems which may cooperate with each other and aid humans. With the increasing amount of software devoted to industrial automation and process control, this capability is becoming more important than ever. Explanations include expressing rationales, characterizing strengths and weaknesses, and projecting their behavior into the future.

Ontologies could play a significant role in explanations since smart engineering systems must represent the conceptual framework that supports explanation. Such explanations would include terms for domain and natural world concepts, relations, and activities. Some version of natural language may be used to describe states and actions in terms that people easily understand, as well as the conceptual structures within which dialog, plans and actions take place.

A benefit of the use of ontologies in support of explanations is the potential for improving interoperability between systems that otherwise would not have a common framework for interoperation. The danger is that current efforts for explainability will be brittle as well as siloed, which will produce a large variety of incompatible explanation techniques that individually satisfy the requirement of providing explanations but which are of little use when explainable subsystems are integrated into large scale systems.

In the usual sense, Ontologies are designed knowledge artifacts but exist in computational (operational) environments which allow reasoning and so should also include the ability to reason about/explain what they know and how they have reasoned with this knowledge. They should be able to express the rationale for the selected use of the relevant parts of an ontology or suite of ontologies; explain the strengths and weaknesses of the chosen ontology; and, when the ontology is in use, explain data that conforms to the ontology.

3 Explainable Artificial Intelligence

Since the 1950s, when the term Artificial Intelligence was coined, there has been considerable progress in this area. The 1980s was dominated with the rise of knowledge-based systems, which is also called "the first wave." Advancements in computer hardware facilitated multilayered neural networks, which led to significant improvements in machine learning for certain classes of problems in 2000s. This was the "second wave." Now, we are witnessing the "third wave," which will include a combination of neural networks and knowledge structures; DARPA views the third wave as contextual adaption where "systems construct contextual explanatory models for classes of real world problems." There are other perspective of various AI waves (e.g., Kai-Fu Lee).

The first wave of AI produced several reasoning techniques, such as decision trees, inference networks, Bayesian networks, statistical learning techniques, and the beginnings of neural networks. As the second wave took place, multilayered neural networks (deep learning) with more than one hidden layer produced impressive results in a wide variety of classification problems. There was a tradeoff between performance and explainability. For example, a major problem with neural networks is the lack of transparency. It was more like a black box approach, where the following questions could not be adequately answered:

- Why did you do that?
- Why not something else?

- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Explainable AI defines a field of research, not define a particular "type" of AI. There exist unique notions of 'explanations', making problem or context-independent explainable AI work inappropriate. "Strong" and "Ultrastrong" AI systems (by Michie's definition) demands the integration of symbolic systems and reasoning to provide operationally effective communications of the internal state and operation of an AI. Meaningful ontologies and knowledge bases, and Symbolic AI methods are a fundamental (but right now, mostly ignored) to the most important XAI use-case: when in practice, within some context, a layman must understand, trust, and be responsible for the conclusions an AI draws.

Sargur Srihari made the bold prediction that the next wave of AI will be probabilistic, which he contrasted with the current wave which is statistical. In this wave, explanations can be inferred probabilistically. Srihari called this the Most Probable Explanation (MPE). He has used this technique in his work on forensics which lends itself to a combination of deep learning and probabilistic explanation.

The DARPA XAI Program was started with the goal to address above questions by enhancing AI software to deal with above questions. [6] See Figures 1 and 2. Note that the DARPA XAI Program proposed that ontologies would be part of the solution. The program also specified that counterfactual arguments are required.

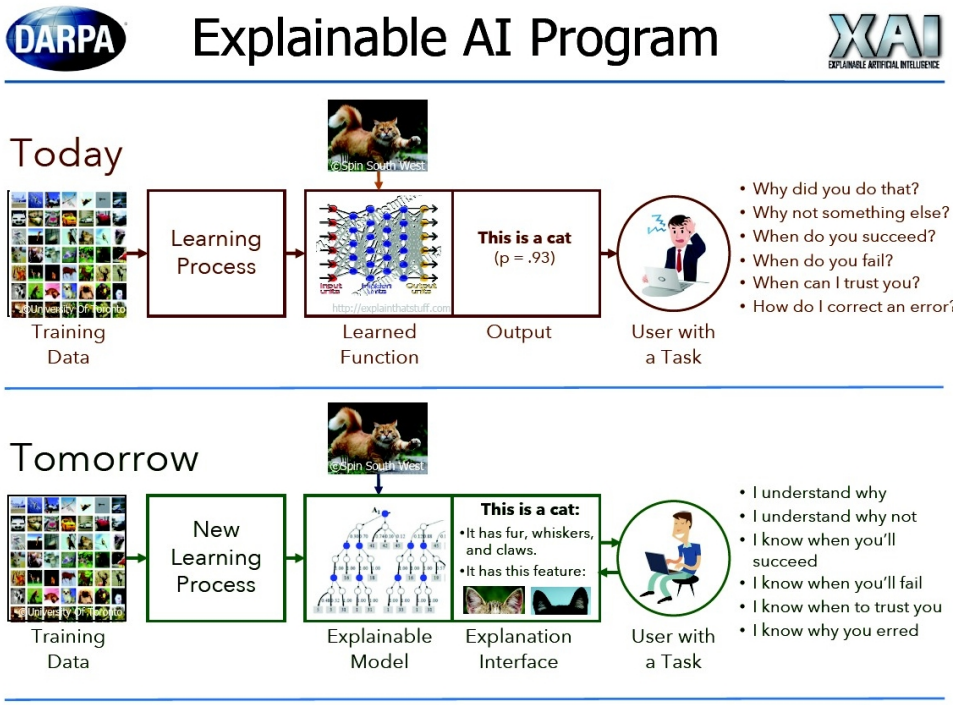

Figure 1: The DARPA XAI Program [6]

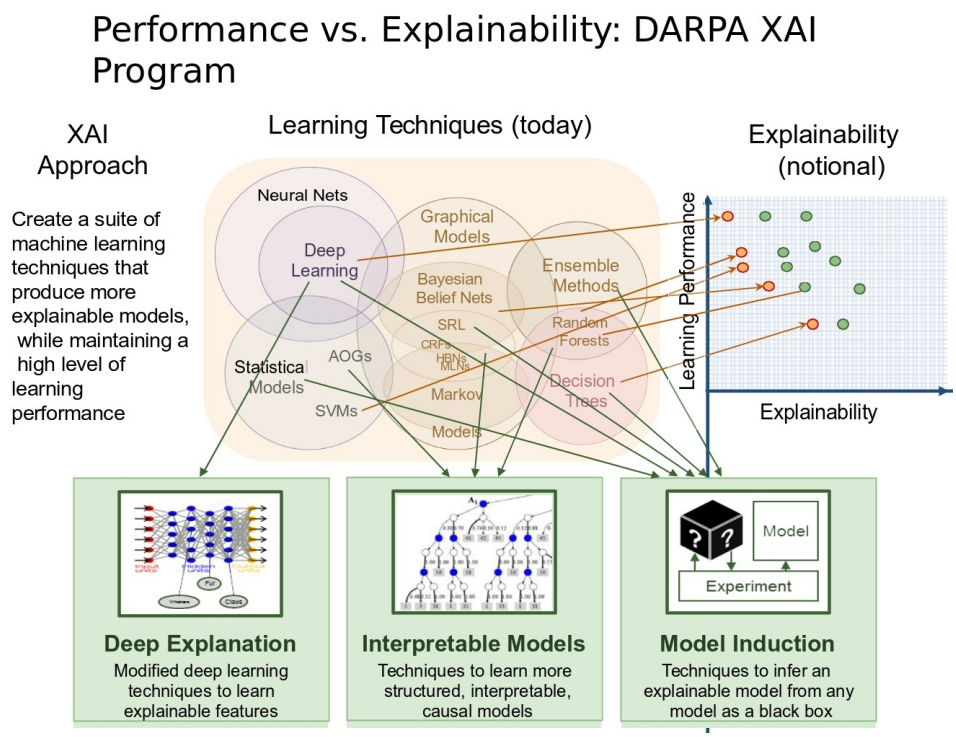

Figure 2: Performance vs Explainability [6]

4 Commonsense Reasoning and Knowledge

There is a long history showing the relevance of commonsense knowledge and reasoning to explanation. Certainly AI founders, such as John McCarthy, believed so and argued that a major long-term goal of AI should include endowing computers with standard commonsense reasoning capabilities. In "Programs with Common Sense" [8] described 3 ways for AI to proceed:

1. imitate the human central nervous system,
2. study human cognition or
3. "understand the common sense world in which people achieve their goals."

Another related goal has been to endow AI systems with natural language (NL) understanding and production. It is easy to see that a system with both Commonsense reasoning and knowledge (CSK) and NL facilities would be able to provide smart advice as well as explanation of this advice. We see the relation of this in the early conceptualization of a smart advice taker system from McCarthy's work that would have causal knowledge available to it for: "a fairly wide class of immediate logical consequences of anything it is told and its previous knowledge." McCarthy further noted that this useful property, if designed well, would be expected to have much in common with what makes us describe certain humans as "having common sense." He went on to use this idea to define CSK - "We shall therefore say that a program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows." [8]

One issue is to address commonsense and explanations for Deep Learning (DL) models. Research is exploring how commonsense could assist in addressing some challenges in DL. The challenges include brittleness of DL models against various adversarial changes to input, generalization to unseen situations when faced with limited training data, and ignoring the overall context while highlighting subtle patterns. Preliminary work suggests that commonsense knowledge and reasoning can compensate for limited training data and make it easier to generate explanations, given that the commonsense is available in an easily consumable representation.

There is evidence that state-change commonsense knowledge can be used for limited training data for procedural text to help prediction. Commonsense aware DL models makes more sensible predictions despite limited training data using the prior knowledge of state-changes e.g., it is unlikely that a ball gets destroyed in a basketball game scenario. The model injects commonsense at decoding phase by re-scoring the search space such that the probability mass is driven away from unlikely situations. This results in much better performance as a result of adding commonsense.

Explanations in the DL context under discussion can be discussed along three dimensions. In the recent years, datasets have fueled research in DL. In both NLP and computer vision community datasets containing explanations have gained interest. This has led to models that provide explanation in the form of attention, natural language sentences and structures such as scene graphs and state-change matrices. Evaluating explanation has remained a challenge, and some evaluation metrics include string matching (exact and METEOR), human based, automated, and, finally, learning to evaluate.

While commonsense is an important asset for DL models, its logical representation such as microtheories has not been successfully employed. Instead, tuples or graphs comprising of natural language nodes has shown some promise, but these face the problem of string matching i.e., linguistic variations. More recent work on supplying commonsense in the form of adversarial examples, or in the form of unstructured paragraphs or sentences has been gaining attention recently.

Some general recurring questions that are worth considering include:

1. How can we leverage the best of the two most common approaches to achieving commonsense? formal representations of commonsense knowledge (e.g. encoded in an ontology's content as in Cyc or Pat Hayes' Ontology of Liquids) vs. strategies for commonsense reasoning (e.g. default reasoning, prototypes, uncertainty quantification, etc.)
2. How to best inject commonsense knowledge into machine learning approaches? Some progress on learning using taxonomic labels, but just scratches the surface
3. How to bridge formal knowledge representations (formal concepts and relations as axiomatized in logic) and representations of language use (e.g. Wordnet)

Commonsense knowledge and reasoning could assist in addressing some challenges in DL as well as explanation and in turn interest in these can turn to CSK and reasoning for assistance.

5 The Role of Narrative

Narrative may be regarded as being complementary to commonsense reasoning and knowledge. While CSK provides the underlying reasoning and knowledge necessary for explanation, the topic of narrative

covers 1) the use of knowledge elements in forming stories as sequences of events, and 2) the presentation of these stories in dialogue.

Narrative is a strategy for sense-making. The Project Narrative at Ohio State University states, "Narrative theory starts from the assumption that narrative is a basic human strategy for coming to terms with fundamental elements of our experience, such as time, process, and change, and it proceeds from this assumption to study the distinctive nature of narrative and its various structures, elements, uses, and effects."[10]

As already noted above in the very definition of explanation, an explanation narrative is an interactive conversation. People leverage their partial understandings of the causal structure of the world "by knowing how to access additional explanatory knowledge in other minds and by being particularly adept at using situational support to build explanations on the fly in real time."[7]

Results from social science regarding how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence. [10, 9] Key insights include:

- Explanations of social behavior rely on theories of an actor's beliefs, desires, intentions and traits, all of which must be considered when generating explanation narratives.
- In addition to causal explanations, i.e., "Why P," explanation narrative also requires contrastive (counterfactual) answers, i.e., "Why P rather than Q?"
- The selection of such causes tends not to be comprehensive but rather identifies a few causes considered to be adequate for the explanation.
- Similarly, one must be selective in the choices of the counterfactual simulations. The use of counterfactual explanation in narrative is a common part of conversation. Automated explanation must be capable of arguing both in favor of an answer as well as against proposed alternative answers.
- As interactive conversations, explanation narratives must abide by rules and maxims of discourse, such as those identified by Grice [4] as follows:
  - The Maxim of Quality
  - The Maxim of Quantity
  - The Maxim of Relation
  - The Maxim of Manner

Relating the conversational evaluation criteria above to the relevance, usefulness and trustworthiness criteria previously noted for explanations can yield additional insights. For example, the Maxim of Quantity suggests that an explanation should be as extensive as necessary but no more so. Surprisingly, the trustworthiness criterion is another one that also has this property. If humans trust a system too much, then they will be unprepared for situations in which the system exhibits anomalous or dangerous behavior. [11]

Many techniques have been developed for generating NL from knowledge and reasoning processes. It is commonly assumed that proofs are the "gold standard" for explanation. However, in practice, proofs are not adequate as explanations. Surprisingly, proofs in the Mathematics literature are far from being complete and rigorous. Their purpose is to convince other mathematicians in the same specialty (which could be quite small) that the theorem is probably true. In most cases, one could generate a complete and rigorous proof from the published argument, but it is very difficult to accomplish and rarely done. In some cases, the claimed theorem turns out to be incorrect in spite of having a thoroughly reviewed and

published "proof." Alternatively, Baclawski has proposed that by taking explanation as the goal, one can develop fully complete and rigorous proofs that can also be readable narratives. [1]

Tiddi[12] studied theories and models of explanation in a variety of disciplines to identify the minimal 'story' elements that define an explanation. The resulting Explanation Ontology Design Pattern is shown in Figure 3 In this pattern, an explanation requires:

- An explanans (the explanation)
- An explanandum (what is being explained)
- A context (or situation) that relates the two
- Some theory or "a description that represents a set of assumptions for describing something, usually general. Scientific, philosophical, and common-sense theories can be included here."
- An agent producing the explanation

The theory element is to be taken as including scientific, philosophical, and common-sense theories as well as concepts defined in cognitive science such as laws, human capacities, human experiences, universals. Theory is a specialization of the Description class from the Descriptions & Situations ontology. "In this view, our theory acts as the description that classifies the explanation (corresponding to the situation), which is built upon the context, the event to be explained, and the possible explaining events."[12]

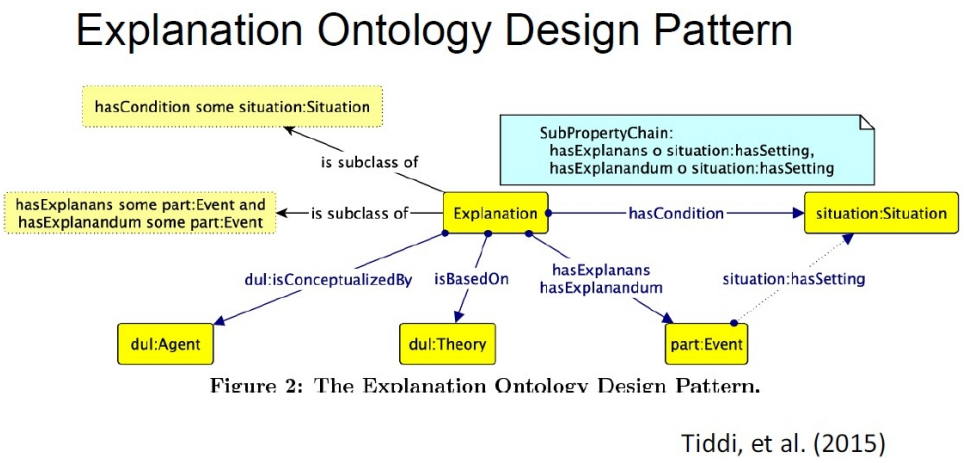

Figure 3: Explanation Design Pattern from [12]

6 Financial Explanations

"Our doubts concerning the near-term feasibility of meeting all the practical challenges facing traditional approaches to explanation do not imply that we assume that a fully intelligent explanation system embodying a passive mode of learning would resolve the problem. On the contrary, even if we were to postulate the existence of such a system, the epistemological objections raised here, as well as in the literatures of education and psychology, would remain unanswered. In short, we favor the design of explanation facilities that acknowledge and exploit the active role played by the learner/user in the process

of meaning making. In response to many of the current and envisioned approaches to explanation, we are inclined to offer the slogan, 'too much instruction – not enough human construction."' [3]

7 Medical Explanations

The health care enterprise involves many different stakeholders – consumers, health care professionals and providers, researchers, and insurers. AI will play an important role in many tasks that these stakeholders undertake. These include: image diagnostics, medical decision making, prior authorization, drug design, nutrition advisor, patient scheduling, etc. During the late 1970s and early 1980s, the pioneering knowledge-based expert systems (KBES) were in the domain of medical diagnosis (e.g., MYCIN, INTERNIST). These systems relied on rule-based inferencing and probabilistic knowledge networks. In the recent past, neural networks (or deep neural networks [DNNs]) are playing a key role in many medical applications that involves image interpretation and diagnosis. Eric Topol's paper and book provide numerous examples of such DNN-based systems. [14, 13] However, the key problem with DNN-based systems is the lack of explainability, which is very important in medical decision making. A combination of DNNs and knowledge networks will form the basis for future AI systems in medicine. Ontologies will play a major role in providing appropriate explanations.

8 Findings

As the summit proceeded, the overlap among the tracks was increasingly evident. Indeed, many presentations used the same or similar slides. The only track that was significantly different was the Financial Explanations track. The Financial domain is much more highly regulated and the rules are very rigorous. However, even the Financial domain has many of the same challenges and issues for explanation.

Returning to the list of questions in Section 1 above, the summit found the following:

- *Can computer-based explanations be made as informative and useful as an explanation from a well-informed person?* There was general agreement that this is possible. However, it was also agreed that it is a difficult problem.
  - *Are there disconnects among researchers, industry and media, or between users and investors with respect to what constitutes an acceptable or successful explanation?*
  - *Is there general ignorance about what is already possible vs what is well beyond current capabilities?*
- *What role do ontologies play?* While ontologies have been proposed for explanation, for example, the Explanation Ontology Design Pattern [12], there does not appear to be a single predominant role that ontologies play in explanation.
  - *How can one integrate ontological and statistical approaches?* It is generally agreed that this is fundamental to explanation, and there have been many promising attempts to achieve this. No single approach has emerged as the standard, and it remains a research problem.
  - *What is required for an ontology to support explanation?* This remains a research problem. While existing ontologies, such as Wordnet, have been very useful, nearly all ontologies today

are relatively shallow, being little more than taxonomies. Furthermore, practitioners commonly feel that existing ontology languages, especially OWL, are not adequate for fully supporting explanation. Unfortunately, there is also no consensus on what additional features or alternative approach is needed.

Other findings:

- Context is important!
- An ontology must itself be explained.
- Counterfactual argument is necessary.
- There are distinct uses of ontology for explanation.
- Explanation must conform to narrative requirements to be acceptable to humans.
- Translations of explanations between narratives and logic should be seamless and unambiguous so as to avoid any confusion for machine interpretability.
- TBD

9 Challenges and Opportunities

- Enrichment of ontologies by incorporating explainability with such factors as dynamic context and situational awareness (also indicated in the Ontology Summit 2018 [2]), commonsense, narratives, provenance and mapping to NLs
- Incorporating commonsense ontologies pertinent to domains of AI solutions
- Comparing different approaches to explainability and differentiating appropriate approaches for specific areas of XAI
- Solutions based on multiple target users and stakeholders
- Track based challenges and opportunities TBD.

References

[1] K. Baclawski. Proof as explanation and narrative, January 2019. Retrieved December 9, 2017 from `http://bit.ly/2RqQJQ5`.

[2] K. Baclawski, M. Bennett, G. Berg-Cross, C. Casanave, D. Fritzsche, J. Ring, T. Schneider, R. Sharma, J. Singer, J. Sowa, R.D. Sriram, A. Westerinen, and D. Whitten. Ontology Summit 2018 Communiqué: Contexts in Context. *Journal of Applied Ontology*, July 2018. DOI: 10.3233/AO-180200.

[3] K. Ford, A. Canas, and J. Coffey. Participatory explanation. In *FLAIRS 93: Sixth Florida Artificial Intelligence Research Symposium*, pages 111–115, Ft. Lauderadale, FL, April 18-21 1993. Retrieved on June 5, 2019 from `http://bit.ly/318aUbX`.

[4] H.P. Grice. Logic and conversation. In P. Cole and J.L. Morgan, editors, *Speech Acts*, pages 41–58. Academic Press, New York, 1975.

[5] D. Gunning. DARPA Explainable Artificial Intelligence: Program Update, 2017. Retrieved on June 4, 2019 from `http://bit.ly/2ITKwfD`.

[6] D. Gunning. DARPA Explainable Artificial Intelligence, 2018. Retrieved on December 3, 2018 from `http://bit.ly/2s9d4pH`.

[7] F.C. Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006.

[8] J. McCarthy. Programs with common sense. In M. Minsky, editor, *Semantic Information Processing*, pages 403–418. MIT Press, 1968. Originally published in 1959.

[9] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 2018.

[10] T. Miller, P. Howe, and L. Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint*, 2017. arXiv:1712.00547.

[11] S. Rodriguez, J. Schaffer, J. O'Donovan, and T. Höllerer. Knowledge complacency and decision support systems. In *IEEE International Inter-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA)*, 2019.

[12] I. Tiddi, M. d'Aquin, and E. Motta. An ontology design pattern to define explanations. In *Proceedings of the 8th International Conference on Knowledge Capture*. ACM, October 2015. Article no. 3.

[13] E. Topol. In *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Hatchett Book Group, New York, 2019.

[14] E. Topol. High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25:44–56, January 2019.