Toward Meaningful Explanations

Kenneth Baclawski[1] , Mike Bennett[2] , Gary Berg-Cross[3],
Todd Schneider[4], Ram D. Sriram[5]

[1]Northeastern University, Boston, MA USA ,
[2]Hypercube Limited, London, UK ,
[3]RDA/US Advisory Group, Troy, NY USA ,
[4]Engineering Semantics, Fairfax, VA USA ,
[5]National Institute of Standards & Technology, Gaithersburg, MD USA

Data is important! It is how we understand the world, and understanding the world is the special interest and purpose of Science. Understanding information that we gather about the world is an important part of the scientific process. However, data that is not correctly interpreted and understood is less than useless, it can actually be misleading or even damaging. So how can scientists, and people in general, understand their data? How can they understand the meaning of their data? If someone does not already understand some data, there should be a mechanism whereby an understanding is possible; in other words, some way to explain the data. This special issue is intended for a wide range of people who are concerned with meaningful explanations, including philosophers, physical scientists, engineers, linguists, social scientists, and many others.

Simply put, an explanation is the answer to the question “Why?” as well as the answers to related questions such as “How?” and “Why not?” and requests for details and evidence for an answer. Accordingly, explanations generally occur within the context of a process, which could be a dialog between persons, between a person and a system, or an agent-to-agent communication process between two systems. It is important to note that explanations are not limited to textual media. Visual media such as diagrams, pictures and videos can also express explanations as well as or even better than text, especially when such media are interactive, thereby fulfilling the requirement that explanations allow for subsequent questions and extended conversation. Explanations also occur in social interactions when clarifying a point, expounding a view, or interpreting behavior. Another important context where explanations are important is the process of developing some kind of system, not necessarily a software system. Such a process requires the developers to make a series of decisions. The explanation for a decision is called its decision rationale.

This special issue is devoted to the subject of what explanations are and what they mean. The inspiration for this special issue is the Ontology Summit that was held in the first half of 2019. This event was concerned with the role of applied ontologies for explaining decisions made by a system. While ontology is the branch of philosophy that deals with the nature of being, applied ontology builds on philosophy, cognitive science, linguistics and logic with the purpose of understanding, clarifying, making explicit and communicating people’s distinctions and assumptions about the nature and structure of the world. Baclawski et al (2019) summarized the findings and challenges that were identified during the Ontology Summit 2019. More specifically, it focused on critical explanation gaps and the role that ontology engineering could play for dealing with these gaps. This special issue expands on the subject of explanations that was introduced by the Ontology Summit 2019.

A brief history of explanations provides some context for this special issue. Among the first known attempts at understanding the why of explanations as explained in (Chatterjee & Dutta, 2014) were those documented among Indian intellectuals and philosophers, beginning with the knowledge collection called the Vedas (dating back to 5000 BCE). This philosophical tradition included notions of context, logic and explanation that are similar to the modern conceptions. For example, there was a notion of syllogism that explicitly incorporated context into the structure of the syllogism. Explanation was also a part of logical inference. More generally, explanation in the form of a dialog between a teacher and a student appears throughout the Vedas (Satprakashananda, 1965; Chennakesavan, 1980).

Greek intellectuals and philosophers subsequently studied the notion of an explanation. 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 via a process akin to modern induction (Shanske, 2006). In the writings of Plato (e.g., Phedus 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 (αι'τία) to explain things. These were from either the thing's matter, form, end, or change-initiator (efficient cause) (Falcon, 2006). 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, namely, how we come to know something and the nature of reality, are parts of explanation. For example, John Stuart Mill provides a deductivist account of explanation as evidenced by these two quotes: "An individual fact is said to be explained, by pointing out its cause, that is by stating the law or laws of causation, of which its production is an instance," and "a law or uniformity of nature is said to be explained, when another law or laws are pointed out, of which that law is but a case, and from which it could be deduced (Mill 1843)."

While explainability has always be a concern of computer systems, the issue has became especially relevant with the success of artificial intelligence (AI) algorithms, such as deep neural networks, whose functioning is too opaque and complex to be understood easily even by those who developed them. This could limit general acceptance of and trust in these algorithms in spite of their advantages and wide range of applicability. Explainable AI (XAI) is an active research area whose goal is to provide AI systems with some degree of explainability. In "Explainable Artificial Intelligence: An Overview," Sargur N. Srihari surveys the field of XAI. Explanations provided by XAI methods take a variety of forms, ranging from traditional feature-based explanations to "heat-map" visualizations, from illustrative examples to probabilistic modeling. Clearly, XAI is an exciting new area at the frontiers of AI.

When computers were developed, one of the earliest questions was whether they might eventually be as intelligent as humans. The field of AI was created not only to investigate this question but also to actually develop systems that achieved it. A fundamental aspect of human intelligence is that we have "common sense," and the study of this aspect of intelligence has been a part of AI from the beginning. AI has also always emphasized the benefits of providing explanations for system reasoning While commonsense knowledge (CSK) and its associated reasoning processes would seem to be useful for explainability, CSK research has, until recently, been more concerned with knowledge representation than with explainability. In "Commonsense and Explanation: Synergy and Challenges in the Era of Deep Learning Systems" by Gary Berg-Cross, the connections between CSK and explanations are discussed, including the challenges and opportunities. The goal is to achieve fluid explanations that are responsive to changing circumstances, based on commonsense knowledge about the world.

The healthcare enterprise involves many different stakeholders – consumers, healthcare professionals and providers, researchers, and insurers. Sources of health related data are highly diverse and have many levels of granularity. As a result of the COVID-19 pandemic, healthcare issues that were previously only discussed by specialists are now part of the everyday discourse of the average individual. In “Applied Ontologies for Global Health Surveillance and Pandemic Intelligence,” Christopher J. O. Baker, Mohammad Sadnan Al Manir, Jon Hael Brenas, Kate Zinszer, and Arash Shaban-Nejad use Malaria surveillance as a use case to highlight the contribution of applied ontologies for enhancing enhanced interoperability, interpretability and explainability. These technologies are relevant for ongoing pandemic preparedness initiatives.

Financial institutions are very complex entities that play many roles and have many kinds of stakeholders, ranging from customers, to regulators, to shareholders, and to the society as a whole. Given these many responsibilities, it is no surprise that financial institutions “have a lot of explaining to do,” as Michael Bennett so deftly begins his article “Financial Industry Explanation” where he presents some of the challenges of providing meaningful explanation in this domain. Explanations are a special case of the more general requirement of accountability which is becoming an issue for many other domains as well. The lessons learned by the financial industry explainability are likely to be valuable for other domains as well.

Ontologies play a significant role in all of the many research projects referenced by papers in this special issue. However, the ontologies for explainability in XAI, commonsense reasoning, health surveillance, and finance do not seem to have much in common with one another. The final paper, “Decision Rationales as Models for Explanations” by Kenneth Baclawski, attempts to weave the various strands of ontologies for explainability together in a single reference ontology by focusing on the observation that the purpose of most of the systems is to make decisions, and that it is the decisions that need to be explained.

Processes today, whether they are based on software or human activities or a combination of them, or whether they use legacy systems or newly developed systems seldom include explainability. In nearly all cases, explanations are neither recorded nor can be easily generated. Unfortunately, explainability cannot simply be added as another module.

Rather it should drive every process from the earliest stages of planning, analysis and design. Explainability requirements must be empirically discovered during these stages (Clancey 2019). Unfortunately, currently there is little sensitivity to the need for explainability and little experience with addressing it. It is hoped that this special issue will assist stakeholders to develop their systems so that they provide meaningful explanations.

References

Baclawski, K., Bennett, M., Berg-Cross, G., Fritzsche, D., Sharma, R., Singer, J., . . . Whitten, D. (2020).

Ontology Summit 2019 Communiqué: Explanation. Applied Ontology. DOI: 10.3233/AO-200226

Chatterjee, S., & Dutta, D. (2014). An Introduction to Indian Philosophy, Eleventh Impression. Rupa Publications.

Chennakesavan, S. (1980). Concept of mind in indian philosophy. Delhi: Motllal Banarsidass.

Clancey, W. (2019). Explainable AI Past, Present, and Future: A Scientific Modeling Approach. Retrieved on April 28, 2019 from http://bit.ly/2Scjvo6

Falcon, A. (2006). Aristotle on causality. Retrieved 16 September 2020 from https://stanford.io/2ZLknqp

Mill, J. (1843). A system of logic. Harper and Brothers.

Satprakashananda, S. (1965). Methods of Knowledge according to Advaita Vedanta. Advaita Ashram.

Shanske, D. (2006). Thucydides and the philosophical origins of history. Cambridge University Press.