Ontology Summit 2019 Communiqué: Explanations

Kenneth Baclawski[1], Mike Bennett[2], Gary Berg-Cross[3], Donna Fritzsche[4], Ravi Sharma[5], Janet Singer[6], John Sowa[7], Ram D. Sriram[8], Mark Underwood[9], David Whitten[10]

[1]Northeastern University, Boston, MA USA, [2]Hypercube Limited, London, UK,
[3]RDA/US Advisory Group, Troy, NY USA, [4]Sherwin-Williams, Cleveland OH, USA,
[5]Senior Enterprise Architect, Elk Grove, CA USA, [6]INCOSE, Scotts Valley, CA USA,
[7]Kyndi, Inc. USA, [8]National Institute of Standards & Technology, Gaithersburg, MD USA,
[9]Krypton Brothers LLC, [10]WorldVistA USA

Abstract

With the increasing amount of software devoted to industrial automation and process control, it is becoming more important than ever for systems to be able to explain their behavior. In some domains, such as financial services, explainability is mandated by law. In spite of this, explanation today is largely handled in an unsystematic manner, if it is handled at all. The decisions of modern artificially intelligent systems, such as those built on deep neural networks, are especially difficult to explain. The goal of the recent Ontology Summit 2019 was concerned with the role of ontologies for explaining the functioning of a system. More specifically, the Ontology 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. The sessions covered explainable artificial intelligence, commonsense reasoning and knowledge, the role of narrative, and explanations in the fields of finance and medicine. The goal of this Communiqué is to foster research and development of approaches to explanations and to drive towards explanation support which can be incorporated into both knowledge engineering processes and ontology design best practices.

Keywords: ontologies; context; inference; commonsense reasoning; explainable AI; narrative; financial explanations; medical explanations

The final publication is available at IOS Press through http://dx.doi.org/10.3233/AO-200226

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 his presentation at the Ontology Summit 2019, Derek Doran asked the question "Okay but Really… What is Explainable AI?" (Doran, 2018). 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. Reasoning: I can conceptualize how inputs are mapped to outputs.

To illustrate the distinctions among the three levels, consider the case of an application for a loan that was denied. The following are examples of the three types of explanation for answering the question "Why was my application for a loan denied?"

1. Interpretation: The account balance covariate in the logistic regression model used to make decisions explains 89.9 % of residual variance.
2. Explanation: The system denies loan applicants with low bank account balances.
3. Reasoning: The system does not want to approve loans to those who do not show evidence of being able to pay them off.

While all three types of explanation are expressed as natural language narratives, there is a significant distinction among them. The first type rigorously expresses features of the decision in terms of a statistical model. It has the most information and even qualifies as a "proof," but fails to relate the decision to the context in which the decision was made. The second type is much better because it is now expressed in terms of the customer context, even if it is not a fully rigorous proof. Yet the second type only explains the function without explaining why that function is being used. The third type is the best since it now explains why that function was used by the bank. The third type uses formal reasoning to infer the rationale from the other types of explanation (i.e., Interpretation and Explanation). What is not shown in the example for the Reasoning type is that it should allow for an interaction with the customer with the goal of achieving customer acceptance of the explanation.

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 is 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 may not be sufficient if it consists of a single generated answer. There should 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 is the subject of the Ontology Summit 2018, the current Summit may be regarded as a continuation of the previous Summit (Baclawski et al., 2018). Indeed, since context includes the answers to the six question words *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 ontologies 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. This Communiqué begins with the background and history of the notion of explanation in Section 2, based on what we learned in the Fall sessions. The three most important areas that were identified in the Fall and studied at greater depth in 2019 were (1) Explainable AI (XAI), presented in Section 3, (2) Commonsense Reasoning and Knowledge (CSRK), presented in Section 4, and (3) Narratives, presented in Section 5. 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 domain, presented in Section 6 and the Medical domain, presented in Section 7, both of which were broadly defined. The findings of the summit are summarized in Section 8. The Communiqué ends with a conclusion in Section 9 and acknowledgments in Section 10.

2 Background

An explanation is the answer to the question “Why?” as well as 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 in 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.

Among the first known attempts at understanding the ‘why’ of explanations 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.

Later in the classical Greek period, Aristotle provided a view of explanation as part of a deductive process using reason to reach conclusions. In the Posterior Analytics from his *Organon*, Aristotle proposed 4 types of causes (αἰτία) to explain things. These were from either the thing's matter (material cause), form (formal cause), end (purpose or final cause), or agent (efficient cause).

The Summit considered not only AI systems that can explain their actions, but also other smart engineering systems which 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 explainable smart systems by representing the conceptual framework that can be used for constructing explanations. An ontology used for explanation would include terms for domain and natural world concepts, relations, and activities.

Software today, whether a legacy system or a newly developed system, most commonly does not include explainability. Unfortunately, explainability cannot simply be added as another module. Rather it should drive the software engineering process from the earliest stages of planning, analysis and design. Explainability needs must be empirically discovered during these stages (Clancey, 2019). Unfortunately, there is little sensitivity to these needs as well as little experience with addressing them in current systems and work practices (Clancey, 2019). One problem is that there are no standards for evaluating the quality of explanations, so there are no objective criteria for determining whether one has adequately implemented explainability (Baclawski, 2019a; Grüninger, 2019). Another problem is that there can be disconnects among the views of the stakeholders. Users generally view explainability as very useful, but investors do not perceive that it adds value (Grosof, 2019). Still other classical software engineering problems arise such as "mission creep" in which requirements are added later in the development process at a time when it is too late to modify the design to accommodate the new requirements, resulting in cost overruns, delays and poor designs (Grosof, 2019). Finally, there is a need for evaluating the quality of explanations. Quality includes not only subjective measures of explainability but also whether the explanations are verifiable and conform to standards.

Perhaps the most all-encompassing reason for developing explainability is the General Data Protection Regulation (GDPR) which states, "The data subject should have the right ... to obtain an explanation of the decision reached ..." (*Recital 71 GDPR*, 2018; Tiddi, 2019). However, the extent to which the GDPR regulations themselves provide a "right to explanation" is heavily debated. Given that explainability is still a nascent research activity, it should not be surprising that the GDPR stops short of mandating this ability immediately. Nevertheless, the intent is to make it easier for users to obtain explanations, especially for fully automated decision making systems.

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 have many of the same undesirable characteristics that have been criticized for AI systems. One of these is fragility (also called brittleness) in the following sense: two perceptively indistinguishable inputs with the same predicted label can be assigned very different interpretations and explanations (Ghorbani, Abid, & Zou, 2019). Another is siloing in the sense that each product employs completely different and incompatible explanation techniques. While such products individually satisfy the requirement of providing explanations, they cannot be consistently integrated into large scale systems.

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 rule-based systems. This period has been called "the first wave" of AI. Advancements in computer hardware facilitated multilayered neural networks, which led to significant improvements in machine learning for certain classes of problems in the 2000s. This was the "second wave." There have been a number of proposals for what will be the dominant focus of the upcoming "third wave," For example, the Defense Advanced Research Projects Agency (DARPA) views the third wave as contextual adaptation where "systems construct contextual explanatory models for classes of real world problems" (Gunning, 2018).

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. During the first wave, it was already recognized that these systems were unable to provide adequate explanations (Jackson, 1986). As Ford et al surmised, "...the inadequate approach to explanation found in most expert systems can be a significant impediment to user acceptance" (Ford, Canas, & Coffey, 1993).

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?

The DARPA XAI Program initially had a goal of addressing the above questions by enhancing AI software to answer them (Gunning, 2018). See Figures 1 and 2. Note that the DARPA XAI Program proposed that ontologies should be part of the solution. The program also required the use of counterfactual arguments.

Sargur Srihari proposed that probabilistic AI (PAI) will be an important part of the third wave of AI. While probability and statistics are related, there are fundamental differences between them. Statistics analyzes observations to understand them, most commonly in the form of a probabilistic model. Probability deals with models that allow one to make predictions and decisions when there is uncertainty. The probabilistic models can be postulated *a priori* or estimated *a posteriori* using statistics. PAI focuses on generative models in the form of probability distributions that can be queried. PAI uses a variety of algorithms and architectures for knowledge representation, inference and learning. It uses Probabilistic Graphical Models (PGMs) to find the most probable explanation (MPE), also called the maximum a posteriori probability assignment. As a result, PAI is more suited to explainability than the techniques employed in the first two waves of AI. Sargur Srihari has used PGMs and MPE in his work on forensics which lends itself to a combination of deep learning and probabilistic explanation (Srihari, 2019).

Explainable AI defines a field of research, not a particular "type" of AI. There exist unique notions of 'explanations', making problem- or context-independent explainable AI work inappropriate. For AI systems to be understandable, symbolic systems and reasoning should be integrated to provide operationally effective communications of the internal state and operation of the system (Michie, 1988). Meaningful ontologies and knowledge bases could play a significant role in ensuring that stakeholders understand and trust the conclusions that an AI system draws.

4 Commonsense Reasoning and Knowledge

There is a long history showing the relevance of commonsense reasoning and knowledge (CSRK) to explanation. Certainly many pioneers of modern AI 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", John McCarthy described 3 ways for AI to proceed: (McCarthy, 1968)

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

While these were proposed as alternative approaches to AI, achieving robust AI is likely to require a cross-disciplinary approach that includes all three.

A related goal has been to endow AI systems with natural language (NL) understanding and production. An early example of this is McCarthy's conceptualization of a smart advice taker that would have access to: "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 CSRK – "We shall therefore say that a program

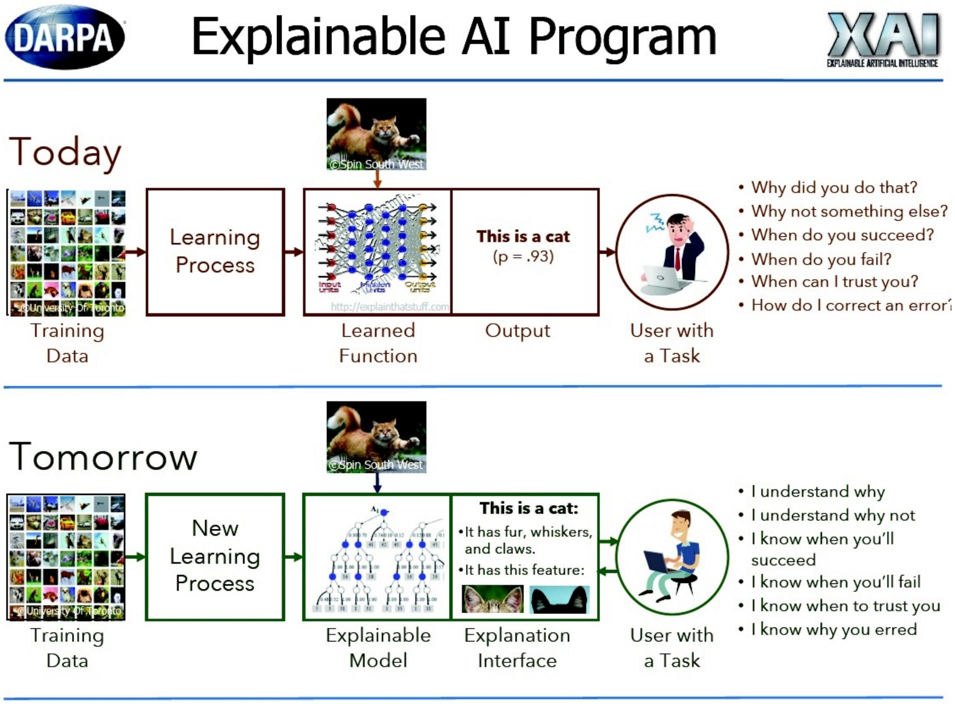
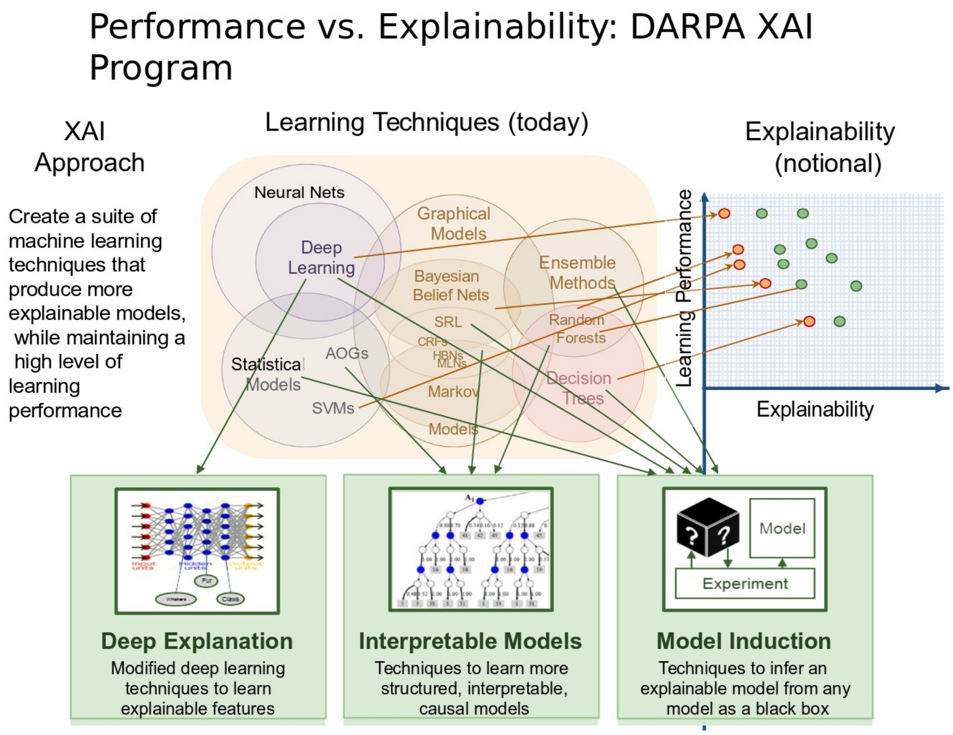


Figure 1: The DARPA XAI Program (Gunning, 2018)

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" (McCarthy, 1968).

Unfortunately, the early attempts to realize McCarthy's conceptualization proved to be brittle and did not capture deep knowledge. These attempts demonstrated that the nature and scale of both explanation and CSRK were difficult. The amount of everyday knowledge needed for common tasks was considerably larger than expected. Indeed, understanding even the simplest children's story, which is a feat that children master with what seems a natural process, is still a formidable problem. The Cyc project attempted to encode a broad range of human commonsense knowledge as a step toward understanding text, which would bootstrap further learning. But this type of progress has been slow, and some believe that today this problem of scale can be addressed in new ways such as ML. But ML methods do not include a built-in way to provide machine generated explanations of what they "know." For this reason, research is exploring how common sense could assist in solving this problem. Particular challenges include addressing the known brittleness of ML models given various adversarial changes to input, and generalization to unseen situations when faced with limited training data. Preliminary work suggests that CSRK could help compensate for limited training data and make it easier to generate explanations, assuming that common sense


Performance vs. Explainability: DARPA XAI Program

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance

Deep Explanation: Modified deep learning techniques to learn explainable features

Interpretable Models: Techniques to learn more structured, interpretable, causal models

Model Induction: Techniques to infer an explainable model from any model as a black box

Figure 2: Performance vs Explainability (Gunning, 2018)

is available in an easily available form.

There is evidence, for example, that state-change commonsense knowledge can be used to compensate for limited training data when the text is describing a procedure or sequence of discrete, inter-related events. Common sense aware ML models make more sensible predictions despite limited training data when they can use the prior knowledge of state-changes. For example, it is unlikely that a ball would be destroyed in a basketball game scenario. The model injects common sense at the decoding phase by re-scoring the search space such that the probability mass is driven away from unlikely situations. Adding common sense results in much better and more robust performance.

The conclusion is that CSRK can assist in addressing some of the challenges of explainability in general and of explainability of AI systems in particular.

5 The Role of Narrative

Narrative may be regarded as being complementary to commonsense reasoning and knowledge. While CSRK 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 dialog.

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” (Miller, Howe, & Sonenberg, 2017).

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” (Keil, 2006).

Results from social science regarding how humans explain to each other can serve as a useful starting point for explanation in artificial intelligence (Miller, 2018; Miller et al., 2017). Key insights include:

- Explanations of social behavior rely on theories of an actor’s beliefs, desires, intentions and traits, all of which should be considered when generating explanation narratives.

- In addition to causal explanations, i.e., “Why P,” an explanation narrative should also provide 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, the choices of the counterfactual answers tend also not to be comprehensive but rather need only be adequate. The use of counterfactual explanation in narrative is a common part of conversation. Automated explanation should be capable of arguing both in favor of an answer as well as against proposed alternative answers.

- As interactive conversations, explanation narratives should abide by rules and maxims of discourse, such as those identified by Grice as follows: (Grice, 1975)

    - 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 (Rodriguez, Schaffer, O’Donovan, & Höllerer, 2019).

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, as we remarked in Section 1 above. In most cases, one could generate a complete and rigorous proof from the published argument, but it is very difficult to accomplish this and it is 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 could also be readable and convincing narratives (Baclawski, 2019b).

Narratives need not be limited to NL. One could also use visual techniques. They are gaining maturity and could be used both for input from, and for output to, the user. Combining NL and visual techniques could be an especially effective approach to explainability.

Tiddi studied theories and models of explanation in a variety of disciplines to identify the minimal ‘story’ elements that define an explanation (Tiddi, d’Aquin, & Motta, 2015). 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 commonsense theories can be included here.”
- An agent producing the explanation

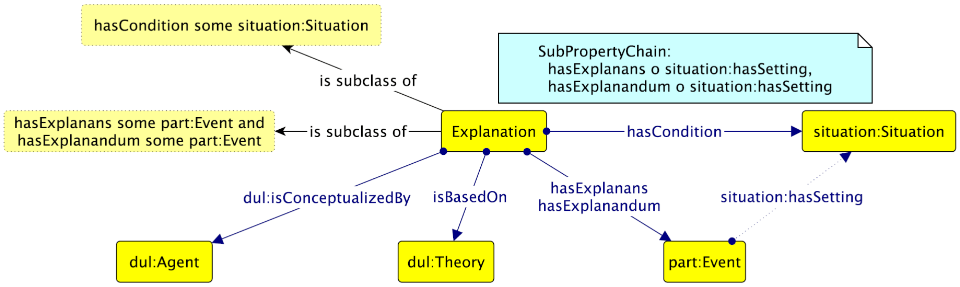


Figure 3: Explanation Ontology Design Pattern from (Tiddi et al., 2015)

Narrative can play a significant role in explanations that are intended for non-specialists and for dialogs. While there is an extensive theory of narration, there are few systems that automatically generate explanations as narratives. Ontologies could be useful for solving this problem.

6 Financial Explanations

The financial services industry is diverse, international and huge. Depending on how it is defined, it globally represents a \$10 to \$20 trillion industry (in U.S. currency). It is also rapidly evolving, with new investment vehicles and services constantly being introduced. This industry is highly regulated, and governments are increasingly mandating that explanations for decisions must be provided to customers (which includes both individuals and companies), to financial services partners, and to government regulators.

Consider the example of the credit industry, a significant subdomain of financial services. In the U.S., the Fair Credit Reporting Act (FCRA) mandates that credit decisions must be explained. For example, a customer whose application for a credit card was denied has the right to know why. The credit card company might respond to a request for an explanation with "Because you didn't meet our criteria." Needless to say, the customer is likely to ask which criteria were not met, and a dialog will then be necessary.

The following are some of many challenges involved in providing explanations in the credit industry in particular, and the financial services industry in general:

- Context
  - *Who* and *Where*: Must comply with Know Your Customer (KYC) regulations. These regulations are intended to prevent illegal activities such as money laundering and bribery.
  - *Who*: Privacy must be respected.
  - *Who*: Certain individuals are protected, such as veterans.
  - *When*: There are temporal constraints, which are complicated by reporting lag times.
  - *What*: Product-specific credit scenarios affect the explanations.
- Data fusion and coordination
  - Must be coordinated with credit reporting agencies (third party).
  - Credit cards can have multiple authorized persons.
  - Rule-based decision support software should be explainable but could have been supplied by a different company.
- Narrative and dialog
  - Most human understanding is contextual, and this needs to be taken into account in presenting explanatory material.
  - The presentation should be easily understandable to humans.
  - Follow up inquiries for more detailed explanations should be supported.
- Explanations of the reasoning process

- When decision making involves complex reasoning processes, the reasoning process itself may need to be explained.
- Higher-order concepts may need to be explained, such as rules and mathematical constructs.

One approach to the challenges listed above would be for the financial services industry to develop a common language. Although this industry is highly regulated, it has no common industry language. The U.S. Treasury Office of Financial Research stated, "The lack of a common industry language is a billion dollar problem." One reason why it is so difficult to develop such a language is that regulations are very large and written in unstructured, natural language text. A single U.S. regulation, the Federal Reserve System Final Rule 12 CFR Part 223, consists of 143 pages of text, and the summary alone is 19 pages long. Another complication is that financial regulations have long reference chains to other regulations.

Nevertheless, progress is being made with the development of the Financial Industry Business Ontology (FIBO) (Bennett, 2019). The FIBO ontology has the potential to contribute to a number of challenges from the list above. As a common language, it would assist in data fusion and coordination. As a formal language, it can be used for explaining reasoning processes. Finally, since FIBO uses natural language terminology, narratives could be formulated to deal with human understanding and dialog.

7 Medical Explanations

The health care enterprise involves many different stakeholders – consumers, health care professionals and providers, researchers, and insurers. AI is likely to 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-I). These systems relied on rule-based inferencing and probabilistic knowledge networks. Unfortunately, as noted in Section 3, most of these systems did not provide adequate explanations of their decisions. In 1993, Ford et al surmised, "...the inadequate approach to explanation found in most expert systems can be a significant impediment to user acceptance" (Ford et al., 1993).

Currently, 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 (Topol, 2019a, 2019b). However, the key problem with DNN-based systems is the lack of explainability, which is very important for many reasons. Medical AI systems will be accepted only if they are trusted by clinicians and other stakeholders, and explainability is essential for this, since patient lives depend on medical decisions. There are also laws and regulations giving patients the right to explanations about their diagnoses and treatments. Explanations are more significant in some medical fields, such as mental health (Kursuncu, Gaur, Thirunarayan, & Sheth, 2019).

A combination of DNNs and knowledge graphs (ontologies) could potentially form the basis for future AI systems in medicine. Achieving this requires solving several problems as discussed in the Introduction above. One must first be able to interpret the DNN. In other words, identify why an input goes to an output. Next one must integrate the interpretation of the DNN with the knowledge graph so that one can trace the reason for the input producing the output (Kursuncu et al., 2019). One must then understand the concepts in the knowledge graph. Finally, a human readable narrative should be generated as discussed in Section 5.

While there is general agreement on the steps required for the generation of explanations given above, there are several possibilities for how knowledge graphs could be used in the machine learning architecture. While there are many ML algorithms, most are variations on supervised learning, i.e., first train the model on labeled training data, then apply the model to new unlabeled data. Based on this architecture, there are at least three ways in which knowledge graphs could be used for explainability:

1. Knowledge enhancement *before* a model is trained

2. Knowledge harnessing *after* the model is trained

3. Knowledge infusing *while* the model is employed

Kursuncu et al. have developed and evaluated all three of these roles for knowledge graphs in the mental health domain using the same ontology for each role (Kursuncu et al., 2019).

Other forms of learning, such as semi-supervised and unsupervised learning may also have multiple roles for ontologies. The Global Health Intelligence project, which involved data integration and interoperability for malaria monitoring and evaluation in sub-Saharan Africa, integrated and analyzed large volumes of structured and unstructured data, and processed the data to find patterns. The project also produced explanatory and predictive models (Shaban-Nejad, 2019). In this case, each of the steps, integration, analysis, pattern recognition and model development can make use of ontologies prior to, during and following the step.

8 Findings

We now summarize the main findings of the Summit. The findings were organized into groups that very roughly correspond to the phases of a system development project. Ontologies can play important roles in achieving explainability throughout all development phases. We begin in Section 8.1 with the reasons *why* a system should be able to explain its functioning. This is followed by Section 8.2 which proposes *what* is necessary for explainability. Section 8.3 discusses challenges for *how* explainability could be achieved. Each challenge represents an opportunity for research on explainable systems.

8.1 Rationale

As we have seen, systems that can explain themselves are difficult to develop. Accordingly, there must be a compelling rationale for allocating resources to support explainability. In this section we mention a few of the reasons why a system should be explainable.

- **Legal requirements.** As discussed in Section 2, explainability is strongly recommended by regulations such as the General Data Protection Regulation (GDPR) (*Recital 71 GDPR*, 2018; Tiddi, 2019). While explainability is not yet mandated in general, some domains already have legal requirements. Financial services have legal requirements for explainability (Bennett, 2019; Underwood, 2019). Forensics is another domain that effectively requires explainability since forensic reports are intended to be used in legal proceedings (Srihari, 2019). Medical systems often have legal requirements for explainability, as well as guidelines and protocols that must be followed. Explainability is also important for systems to be trusted and accepted by users (Kursuncu et al., 2019).

- **Trust.** Even in domains that do not have requirements for explainability, the need is still compelling. In many domains, AI is still a novelty and explanation may be necessary to gain acceptance and trust by practitioners (Kursuncu et al., 2019; Turano, 2019).

- **Improve robustness.** It is known that AI systems are fragile and can be easily fooled (Ghorbani et al., 2019). Explanation may be able to help to improve robustness and to deal with adversaries (Clancey, 2019; Tandon, 2019).

- **Interoperability.** One of the advantages of ontologies, especially when they have been standardized, is the potential for improving interoperability and coordination among systems. The need for this capability was seen in both financial services and medical systems (Sections 6 and 7).

8.2 Requirements

Having made a case for systems to be explainable, the question then becomes what is required. As discussed in Section 2 above, an explainable system should be able to answer commonsense questions, especially “why” questions, but also “how” and “why not” questions. The questions and answers should be in natural language, should be expressed in an easily understandable narrative style, and should allow for interactivity and following up (Berg-Cross & Hahmann, 2018; Clancey, 2019; Sowa, 2018; Sriram & Sharma, 2018).

- **Interdisciplinary.** A common theme during the Summit sessions was that explainability is very likely to require cross-disciplinary teams from computer science, cognitive science, linguistics, social science, and biological sciences. Within computer science, techniques from many fields must be combined, including classical logical (symbolic) AI, statistical (sub-symbolic) AI, NL processing, NL generation, ontologies, and CSRK (Sriram & Sharma, 2018). Combining such disparate techniques so that provenance remains traceable is a substantial challenge.

- **Ontologies.** There was a general consensus among the invited speakers and participants that ontology can be useful for explainability. Indeed, most claimed that ontology was essential for explainability. While ontologies have been proposed for explanation, for example, the Explanation Ontology Design Pattern (Tiddi et al., 2015). While such patterns are useful, there does not appear to be a single predominant role that ontologies play in explanation. An ontology could be in the "background" for purposes such as commonsense reasoning (Berg-Cross & Hahmann, 2018). On the other hand, an ontology could also be more explicitly visible to end users as in financial services (Bennett, 2019; Underwood, 2019). When this is the case, the ontology must itself be explainable (Bennett, 2019). When an ML algorithm is being employed, there are many roles that an ontology could play as discussed in Section 7 (Kursuncu et al., 2019; Shaban-Nejad, 2019). Some of these are background roles, while others are explicit and require the ontology to be explained. In general, there is a need for more research on ontology design patterns for explanations.

- **Context.** It was also generally agreed that context is important (Bennett, 2019; Berg-Cross & Hahmann, 2018; Clancey, 2019; Sriram & Sharma, 2018; Underwood, 2019). Indeed, as noted in Section 1, context is part of the definition of explanation. Moreover, explanation may be regarded as being a part of the context as well as depending on it. Specifically, explainability requires a "user model" of interests and knowledge in order to have meaningful dialogs (Clancey, 2019). A more subtle aspect of the user model is the user's language. This is not just the natural language of the user but also the domain-specific jargon of the user's community.

- **Quality.** Finally, there is a need for evaluating the quality of explanations (Baclawski, 2019a; Grüninger, 2019). Quality includes not only subjective measures of explainability (such as user satisfaction surveys) but also whether the explanations are verifiable and conform to standards.

8.3 Challenges and Opportunities

Explainability is a difficult research problem, but there was general agreement that it will be possible to develop systems that can produce explanations that are as informative and useful as an explanation from a well-informed person. That said, there are some serious challenges that represent opportunities for research.

- As discussed in Section 2, explainability should be an integral part of software development, even to the extent of driving the development process. Adding explainability later in the process can be less adequate and more expensive than including it early in the process. This is a difficult problem because stakeholders have different attitudes toward support for explanations.

- By the definition of explanation, it depends on context, both the system context and the user context. Consequently, explainability will require solutions to many of the problems that were found in the Ontology Summit 2018 (Baclawski et al., 2018).

- In the financial domain, regulations are very large, unstructured natural language documents with long reference chains. Formalizing such regulations correctly and efficiently is a barrier to automating explanations in this domain. A further complication is that the constructs need not be first-order, making it difficult to apply existing reasoning tools (Bennett, 2019).

- As noted in Section 4, commonsense reasoning is a key enabler for a robust AI explanation system (Berg-Cross & Hahmann, 2018; Tandon, 2019; Tiddi, 2019). Unfortunately, while great progress has been made in this area, it remains a research problem. Consequently, common sense is "perhaps the most significant barrier" between the focus of AI applications today, and the human-like systems we dream of (Berg-Cross & Hahmann, 2018).

- One barrier for developing explainable systems that interact with humans on everyday tasks is the lack of a common framework for modeling such tasks. An ontology that we all share for representing and reasoning about the physical world could contribute to solving this problem (Grüninger, 2019).

- While mathematical proofs can be useful for explainability, automated proofs are not easy to understand by a non-specialist (Doran, 2018; Baclawski, 2019a; Berg-Cross & Hahmann, 2018). There is a need to apply linguistic and narrative theories to mathematical proofs to produce more easily understandable explanations. However, expressions of proofs as narratives should be seamless and unambiguous so as to avoid any confusion for machine interpretability.

- Sharing and reusing insights from the social and cognitive sciences is a key enabler for developing explainable systems (Tiddi, 2019).

- The Most Probable Explanation is a very promising approach for explainability (Srihari, 2019).

- Visual techniques are a promising approach to explanation, but automatically generating explanatory visualizations that are integrated with narratives is challenging.

9 Conclusion

Automating explanations of complex systems today is largely handled in an unsystematic manner, if it is handled at all. Modern AI systems are especially difficult to explain, but they are not the only systems that lack adequate explainability. We found that ontologies and context sensitivity can be important parts of an effective solution to explainability. This conclusion was apparent in all of the Summit tracks, both in the general areas of commonsense reasoning and narrative and in the more specific domains of explainable AI, financial services and medicine. While we listed many potential approaches that are actively being pursued, much more research and development is needed for systems to be adequately explainable.

10 Acknowledgments

Certain commercial software systems are identified in this paper. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology (NIST) or by the organizations of the authors or the endorsers of this Communiqué; nor does it imply that the products identified are necessarily the best available for the purpose. Further, any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of NIST or any other supporting U.S. government or corporate organizations.

We wish to acknowledge the support of the ontology community, especially the invited speakers and participants who contributed to the Ontology Summit. There were 23 invited speakers, some of whom gave presentations at more than one session. The complete list of sessions, speakers, and links to presentation slides and video recordings is available at `http://bit.ly/2WTduUB`. We also wish to acknowledge feedback from many other individuals, including Dieter Gawlick, Paul Sonderegger, Anna Chystiakova, Kenny Gross, Zhen Hua Liu, Azamat Abdoullaev, and Elham Tabassi.

References

Baclawski, K. (2019a, January). *Introduction to the Summit Tracks.* Retrieved on June 12, 2019 from `http://bit.ly/2VRt6DV`

Baclawski, K. (2019b, January). *Proof as explanation and narrative.* Retrieved on June 12, 2019 from `http://bit.ly/2RqQJQ5`

Baclawski, K., Bennett, M., Berg-Cross, G., Casanave, C., Fritzsche, D., Ring, J., ... Whitten, D. (2018, July). Ontology Summit 2018 Communiqué: Contexts in Context. *Journal of Applied Ontology*. DOI: 10.3233/AO-180200

Bennett, M. (2019, February). *Ontology Summit 2019 Financial Explanations Track Session 1: Financial Industry Explanations.* Retrieved on April 28, 2019 from `http://bit.ly/2RIViFQ`

Berg-Cross, G., & Hahmann, T. (2018, December). *Overview of Commonsense Knowledge and Explanation.* Retrieved on April 28, 2019 from `http://bit.ly/2Uhbxwh`

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

Doran, D. (2018, November). *Okay but Really... What is Explainable AI? Notions and Conceptualizations of the Field.* Retrieved on April 28, 2019 from `http://bit.ly/2TM2141`

Ford, K., Canas, A., & Coffey, J. (1993, April 18-21). Participatory Explanation. In *Flairs 93: Sixth florida artificial intelligence research symposium* (pp. 111–115). Ft. Lauderadale, FL. Retrieved on June 5, 2019 from http://bit.ly/318aUbX

Ghorbani, A., Abid, A., & Zou, J. (2019). Interpretation of neural networks is fragile. *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*, *33*. arXiv:1710.10547

Grice, H. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Speech acts* (pp. 41–58). New York: Academic Press.

Grosof, B. (2019, January). *An overview of explanation: Concepts, uses, and issues.* Retrieved on April 28, 2019 from http://bit.ly/2R9iD30

Grüninger, M. (2019, January). *Ontologies for the physical turing test.* Retrieved on April 28, 2019 from http://bit.ly/2RiQSFz

Gunning, D. (2018). *DARPA Explainable Artificial Intelligence.* Retrieved on December 3, 2018 from http://bit.ly/2s9d4pH

Jackson, P. (1986). *Introduction to expert systems.* Addison-Wesley, Reading, MA.

Keil, F. (2006). Explanation and understanding. *Annu. Rev. Psychol.*, *57*, 227–254.

Kursuncu, U., Gaur, M., Thirunarayan, K., & Sheth, A. (2019, March). *Explainability of Medical AI through Domain Knowledge.* Retrieved on June 12, 2019 from http://bit.ly/2YuPUe3

McCarthy, J. (1968). Programs with common sense. In M. Minsky (Ed.), *Semantic information processing* (pp. 403–418). MIT Press. Originally published in 1959

Michie, D. (1988). Machine learning in the next five years. In *Proceedings of the third european working session on learning* (pp. 107–122). Pitman.

Miller, T. (2018). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*.

Miller, T., Howe, P., & Sonenberg, L. (2017). Explainable AI: Beware of inmates running the asylum or: How I learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint*. arXiv:1712.00547

Recital 71 GDPR (Tech. Rep.). (2018). European Union. Retrieved June 14, 2019 from http://bit.ly/2KL6LVz

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

Shaban-Nejad, A. (2019, April). *Semantic analytics for global health surveillance.* Retrieved on June 12, 2019 from http://bit.ly/2ZkCzoT

Sowa, J. (2018, September). *Representing and reasoning about contexts using IKL.* Retrieved on May 15, 2018 from http://bit.ly/2xrtkIa and http://bit.ly/2y4h41v

Srihari, S. (2019, April). *Explainable Artificial Intelligence: The Probabilistic Approach.* Retrieved on June 12, 2019 from http://bit.ly/2UlmgES

Sriram, R., & Sharma, R. (2018, November). *Role of ontologies in explaining reasoning: An overview of explainable AI.* Retrieved on April 28, 2019 from http://bit.ly/2TWYzUr

Tandon, N. (2019, January). *Commonsense for Deep Learning.* Retrieved on June 12, 2019 from `http://bit.ly/2XEOcWS`

Tiddi, I. (2019, March). *Building intelligent systems (that can explain).* Retrieved on June 12, 2019 from `http://bit.ly/2SVUD4l`

Tiddi, I., d'Aquin, M., & Motta, E. (2015, October). An ontology design pattern to define explanations. In *Proceedings of the 8th international conference on knowledge capture.* ACM. Article no. 3

Topol, E. (2019a).
In *Deep medicine: How artificial intelligence can make healthcare human again.* Hatchett Book Group, New York.

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

Turano, A. (2019, February). *Review and recommendations from past experience with medical explanation systems.* Retrieved on April 28, 2019 from `http://bit.ly/2WZfMOP`

Underwood, M. (2019, February). *Explanation use cases from retail credit card finance regulatory and service quality drivers.* Retrieved on April 28, 2019 from `http://bit.ly/2WLidE7`

A Full Web Links

http://bit.ly/2KL6LVz	https://www.privacy-regulation.eu/en/r71.htm
http://bit.ly/2R9iD30	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Commonsense/Explanation--BenjaminGrosof_20190123.pdf
http://bit.ly/2RIViFQ	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Financial/Finance+Track+Session+1+-+MikeBennett.pptx
http://bit.ly/2RiQSFz	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Commonsense/summit-physical-turing.pdf
http://bit.ly/2RqQJQ5	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Narrative/ProofAsExplanationAndNarrative--KenBaclawski_20190130.pdf
http://bit.ly/2SVUD4l	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Narrative/BuildingIntelligentSystemsThatCanExplain--IlariaTiddi_20190313.pdf
http://bit.ly/2Scjvo6	https://s3.amazonaws.com/ontologforum/OntologySummit2019/XAI/ExplainableAI--WilliamClancey_20190220.pdf
http://bit.ly/2TM2141	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Overview/Summit-Overview-Session-2-DerekDoran_20181128.pdf
http://bit.ly/2TWYzUr	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Overview/Explainable-AI-and-Ontology--RaviSharma-RamDSriram_20181128.pptx
http://bit.ly/2Uhbxwh	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Overview/IntroductionToCommonsenseAndExplanations--GaryBergCross-TorstenHahmann_20181205.pdf
http://bit.ly/2UlmgES	https://s3.amazonaws.com/ontologforum/OntologySummit2019/XAI/XAI--SargurSrihari_20190410.pdf
http://bit.ly/2VRt6DV	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Introduction/IntroductionToTracks--KenBaclawski_20190116.pdf
http://bit.ly/2WLidE7	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Financial/Explanation+Use+Cases+from+Retail+Credit+Card+Finance.pdf
http://bit.ly/2WTduUB	http://ontologforum.org/index.php/OntologySummit2019#Structure_and_Discourse
http://bit.ly/2WZfMOP	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Medical/ExplainableMedicalAI--AugieTurano_20190213.pptx
http://bit.ly/2XEOcWS	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Commonsense/CommonsenseForDeepLearning--NiketTandon_20190306.pdf
http://bit.ly/2YuPUe3	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Medical/MedicalExplanation--UgurKursuncu-ManasGaur_20190327.mp4
http://bit.ly/2ZkCzoT	https://s3.amazonaws.com/ontologforum/OntologySummit2019/Medical/SemanticAnalyticsForGlobalHealthSurveillance--ArashShabanNejad_20190417.pptx
http://bit.ly/2s9d4pH	https://www.darpa.mil/program/explainable-artificial-intelligence
http://bit.ly/2xrtkIa	http://ontologforum.org/index.php/ConferenceCall_2017_09_20
http://bit.ly/2y4h41v	http://ontologforum.org/index.php/ConferenceCall_2017_09_27
http://bit.ly/318aUbX	https://www.ihmc.us/users/acanas/Publications/ParticipatoryExplanation/ParticipatoryExplanation.htm