Commonsense and Explanation: Synergy and Challenges in the Era of Deep Learning Systems

Gary Berg-Cross

Ontology Forum Board Member

Abstract

This article builds on some of the research ideas discussed in the commonsense reasoning and knowledge track as part of the Ontology Summit 2019 on explanations. As discussed there, research on intelligent systems has long emphasized the benefits of providing explanations for system reasoning, although approaches to an explanation function have evolved over time. While system-provided explanations like common-sense knowledge (CSK) and associated reasoning (CSKR) each go back to the early days of artificial intelligence (AI) systems, they became somewhat independent research areas for much of their later history. This was in part because explanations in early AI efforts were technical in nature centering on how faithfully a system describes the reasoning and heuristic steps employed. Another factor was the difficulty of building adequate bases of CSK for reasoning. Although early AI notionally recognized that as part of intelligent systems explanations should make commonly understood sense, this was not a sustained priority in later work. Instead CSK research engaged more on issues of adequate knowledge representation, how to acquire a base of CSK and the diversity of ontologies needed to support CSK. While these are not finished research areas, they now provide useful guidance to support a current interest in the role of CSK explanations motivated by new challenges and opportunities. These include the rapidly expanding space of heterogeneous and richly interconnected data along with diverse sub-symbolic (deep learning) intelligent system applications. New AI approaches include useful, but only partially understood results, from machine learning (ML) and deep neural net (DNN) approaches. The complexity of these approaches, which includes use of patchy and inconsistent information available online, prompts a renewed desire to have systems explain their decisions and processing in deep, flexible, defendable and understandable ways. Recent work has promoted the development of AI systems using ML-based models with a range of explanatory capabilities for generated decisions. Common sense concepts now play a role in providing better performance and a range of more easily understood explanations for end users.

Taken as a whole, the cumulative lesson of decades of research is that fluid explanations, responsive to changing circumstances require knowledge about the world and that explanations are intimately connected to both common-sense reasoning and background knowledge such as captured in formal ontologies, but also informally understood in text (Davis and Marcus, 2015). The combination of information and its context extracted from a range of sources and organized and represented formally provides a base, not only for intelligent system performance, but also for background knowledge needed for flexible and deep explanations. In practice there seem to be many views of satisfactory explanations but that CSK and reasoning plays varying, but useful roles in each of these.

Among the remaining challenges are those of developing an adequate base of CSK, an adequate approach to situational and contextual understanding, how to use deep learning in dynamic situations, the need to keep humans in the loop and the need for a common enhanced ontology engineering practice addressing both explanation and CSK.

Introduction

THE 2019 ONTOLOGY SUMMIT on Explanation (Ontology Summit, 2019) provided an opportunity to look at various approaches of intelligent/smart systems [i]from a number of perspectives including that of commonsense knowledge (CSK) and associated reasoning (CSKR). Commonsense reasoning and knowledge was prominently featured as an early part of Artificial Intelligence (AI) conceptualization, and it was assumed to be important in the development and enhancement of human-like, intelligent systems explanations, which also had a defined role in early AI. Both continue to be considered important parts of intelligent systems and this is not surprising when we consider the centrality of an ability to explain reasoning and what they know by a system whose claim to fame is intelligence itself. Over the past half century of work on intelligent systems, a variety of approaches to explanation have been engineered and deployed and when carefully designed proven useful. On the whole CSKR's role in explanations has been more indirect than direct. It has often been used to provide a perspective on explanatory short comings. However, new ML techniques that construct and represent knowledge using non/sub-symbolic models layer additional requirements for understandable explanation. This in turn provides and opportunities for CSKR to aid in such explanations (Chakraborty, *et al*. 2017).

In the sections that follow I discuss some of the historical relations of explanation and CSKR followed by some of the experience over time of crafting both CSKR and good explanations. A useful way to illustrate the current status of this work is to overview how some explanation applications are built and employed in representative areas. Following this I overview some of the issues and some of the challenges introduced by a consideration of applying CSKR to contemporary AI and ML systems and the recent efforts in the new field of eXplanatory AI (XAI). We conclude with a summary of some preliminary findings, identification of remaining issues and opportunities that might promote and guide future work.

Some Background on the AI Connections of Commonsense and Explanation

In this section I review some of the major developments along the AI path to intelligent systems and why CSKR seems like an important ingredient in the development of intelligent explanation. Note, that this review is not comprehensive, but represents a survey giving the flavor of methods and results that are pertinent to the evolution of explanations and CSKR.

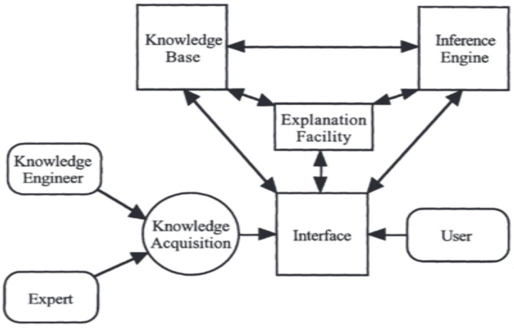
Simply put, fifty years of experience teaches us that only an intelligent system that justifies its actions in terms which make sense so they are readily understandable to the user will be trusted (Cohen *et al*, 2017) . Early AI work showed that rudimentary attempts at explanation provided useful to system engineers and a modicum of user satisfaction if not trusted (Langlotz and Shortliffe, 1984). As a result improvements in explanation have remained a necessary next step in intelligent system evolution for a long time. Interestingly, one sees in the original Turing Test the need for CSKR and explanations each as part of communication to pass the test. These are, of course, common human abilities to live in an ordinary world (Ortiz, 2016). Some examples of CSKR needed for passing a Turing Test or just living in society are illustrative of the range involved and might include the following type of reasoning:

- Taxonomic: Cats are mammals.
- Causality: Eating a peach makes you less hungry.
- Goals: I don’t want to get hot so let’s find shade.
- Spatial: You often find a microwave in the kitchen.
- Functional: You can sit on a chair if tired.
- Planning: To check today’s weather look on a weather application.
- Linguistic: The word “won’t” is the same as “would not”.
- Semantic: Cat and feline have a similar meaning.

Many cognitive abilities that are developed it seems simply in the first years of life provide the commonsense knowledge and reasoning to handle the above list and problems like conservation of objects - if I put my toys in the drawer, they will still be there tomorrow. It has proven much harder to get such an adequate base of knowledge and associated reasoning into computational systems. Early on in this process two ways seem possible to populate such knowledge for an intelligent system. One is by handcrafting in a mass of commonsense knowledge, while another is by letting a system learn from training experience with things like object conservation over time or place. One may also consider some combination of the two, say building in some knowledge and using that to learn more, or letting it learn and correcting errors by adding hard to learn knowledge or by dialog with a user.

Indeed an early AI goal was to endow systems with natural language (NL) understanding and text production, which it worth noting could be used for explanations. It is easy to see that a system with both CSKR and NL facilities would be able to provide smart advice as well as explanations of this advice. We see both in the early conceptualization of a smart advice taker system from McCarthy’s work making causal knowledge available for: “a fairly wide class of immediate logical consequences of anything it is told and its previous knowledge.” McCarthy (1960) 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 commonsense.” John McCarthy believed so and argued that a major long-term goal of AI should include endowing computers with standard commonsense reasoning capabilities.[ii]

While there is a long history showing the relevance of commonsense knowledge and reasoning to explanation in actual practice, going back to the 70s and 80s, AI systems, aka “expert systems”, were not as the founders envisioned. They were less knowledgeable and brittle, based on explicit models of domains implemented using handcrafted production rules encoding useful information about special topics such as diseases. In part because of handcrafting of knowledge rather than the engineering of knowledge systems rule, knowledge was fragmented and opaque and would break down revealing obvious errors. Part of this was due in part to a lack of the robustness available from human-like commonsense which was hard to handcraft or engineer into applications’ supporting knowledge bases. Following an easier development path 1970s era expert systems came, as shown in Figure 1, with a very simple, technical, but not commonsense rich idea, of what was called an "explanation facility." The early implementations used a proof trace of rule firings which provided a purely technical explanation. It did not include what we call justifications for its explanations. Such proofs founded on "Automated Theorem Provers" (Melis, 1998) could provide a map from inputs to outputs and served the needs of system engineers to understand system performance more than providing an explanation to a user[iii].



Example of an Early AI System Architecture
From Medsker, Larry R. Hybrid intelligent systems.
Springer Science & Business Media, 2012.

Figure 1. Simple View of an Early Expert System

But case specific and mathematical based proof planning are not as robust or as reliable as they first seemed to AI developers. This was due to the commonly understood fact that situations being reasoned over were often not adequately represented. Thus, situations and the explanations about them lose some intuitive meanings expected by users (Bundy, 2002). Another problem is that rules in a knowledge base (KB) can change over time and early efforts did not include meta-knowledge to explain why they change. To make sense changes often need explanations.

Along with brittleness and limited utility of traces, part of the weaknesses of rule-based explanatory reasoning, was exposed by Clancey (Clancey, 1983). He found that the AI system called Mycin's had individual rules that play vastly different roles, have different kinds of justifications, and are constructed using different rationales for the ordering and choice of premise clauses in the rules. Since in this rule knowledge isn't made explicit, it can't be used as part of explanations. And there are structural and strategic concepts which lie outside of early AI system rule representations. It was soon realized that these can only be supported by appealing to some deeper and contextual level of (background) knowledge. But commonsense context was seldom "explicitly stated" and thus difficult to engineer.

In searching for solutions the next generation of AI developers used more structured and formal KBs such as frames or semantic net-like ontologies to capture and formalize a fuller range of necessary knowledge. At this time the role of causal-based explanations also helped design more knowledgeable and integrated rather than *ad hoc* expert systems, based on the idea that a system's knowledge should be integrated with performance and adequate to explain its reasoning (Swartout and Smoliam, 1989). Taken together this made the argument that something like ontologies are needed to make explicit structural, strategic, and support-type knowledge. One result was development of large KBs such as in the Cyc project (Lenat and Guha, 1989), a 35-year effort to codify common sense into an integrated, logic-based system. Efforts like Cyc which started up in the 80s represented an effort to avoid problems like system brittleness by providing a degree of common-sense and modular knowledge (Lenat, *et al* 1985). Cyc can provide a response to queries such as: "Can the Earth run a marathon?" In terms of a commonsense explanation we have a "no" because of the knowledge that the Earth is not animate and the role capability needed to run a marathon is detailed by the knowledge in a sports module. Indeed the need for a formal mechanism for specifying a commonsense context had become recognized, and some approach to it, such as Cyc's microtheories arose[iv]. In the 80s Cyc-type knowledge was also seen as important to what was called associate systems. This advance argued that "systems should not only handle tasks automatically, but also actively anticipate the need to perform them....agents are actively trying to classify the user's activities, predict useful sub-tasks and expected future tasks, and, proactively, perform those tasks or at least the sub-tasks that can be performed automatically" (Panton, 2006). All of these abilities were, of course, conceptually useful for explanation, so advances in CSKR, like a Cyc micro-theory could serve a dual role. Much ontological work has followed the spirit of this idea if not the exact program outlined to build large KB such as the Cyc project.

But subsequently, except for a few systems they were rarely applied as part of mainstream systems although the need was often noted (Minsky, 2000). Although some efforts, such as Crowdsourcing common sense training data in Open Mind (Singh, 2002a) are notable, the effort to engineer sufficient CSK for reasoning as well as reasonable explanations has proven difficult. While there are some success as an aid to NLP, where hybrid approaches out perform an NLP tool like BERT (Havasi, 2019), the scale of the problem has been discouraging; for people seem to need a tremendous amount of knowledge of a very diverse variety to understand even the simplest children's story (Singh, 2002b). Research retreated from an ambitious broad CSKR aim and instead pursued special domain knowledge and reasoning that could deal with a more focused class of problem. But these lacked generalization and thus did poorly at almost everything else (Minsky, Marvin L., Push Singh, and Aaron Sloman, 2004).

Despite direct approaches to explanation and problems of formalizing background knowledge, work since the 1990s has included other forms, styles, or meanings of explanation that seemed easier. Because proof isn't always useful and deep background knowledge is hard to formalize another form of documentation, and thus a style of explanation has often been used that involves the provenance or source of some fact or statement (McGuinness, 2003, Moreau, 2010, Darlington, 2013). This arises often when we want an explanation to make clear what the documented source of data is[v].

In contrast AI explanation work in the 90s and early 2000s focused on simpler techniques to make explanations acceptable to novice users rather than using large KBs of CSK which were expensive, time consuming, and hard to build with the tools and limited expertise available. Modest use was made of cognitive learning theory and associated technology[vi] which suggested the need for explanation justification using explicit knowledge of things like conceptual terminology, domain facts, and causal relations to enhance the ability of novice user's understanding (Darlington, 2013). What was more desired was explanations that also aided engineers in modifying systems (*e.g*. knowledge debugging as part of KB development).

CSKR and Explanation in the new era of ML

As noted earlier, it can be costly to acquire an adequate base of CSKR for its own sake as well as leverage it for explanations. And, when acquired, since there are a variety of ways to represent CSKR, from symbolic forms of rules to semantic nets, and logic, the knowledge content becomes heterogeneous and siloed making them difficult to integrate and structure for explanations.

This makes it attractive to consider lighter methods for acquiring knowledge like opportunistic extraction processes from text, including online text and linked data, using AI, ML, or NLP tools. Rapidly advancing ML capabilities have raised the hope of capturing knowledge including masses of CSK in more automated ways that are less resource intensive. There has now been a decade of work to acquire and represent domain knowledge, even some commonsense-like knowledge, using automated extraction and ML processes that acquire models learned from training data. A remaining problem with early work that is still somewhat with us is that a large store of training data is needed because the model must learn anew from scratch each time it learns anything. And this isn't how people work.[vii]

One prominent, illustrative attempt to tackle this problem is the Never-Ending Language Learner (NELL) system which uses a coupled semi-supervised training approach (Mitchell *et al*, 2018). Central to the NELL effort is the idea that we will never truly understand machine or human learning until we can build computer programs that share some similarity with the way humans learn. This does promise the possibility of acquiring a useful set of CSKR along the way. In particular such systems, as discussed by (Mitchell *et al*, 2018), are like people in that with years of diverse, mostly self-supervised experience, they can learn many different types of everyday knowledge or functions and thus information from many contexts. This happens in a staged bootstrapping fashion, where previously learned knowledge in one context enables learning further types of knowledge. It is easy to elaborate on cognitive processes for informed ML (Von Rueden et al, 2019) using ideas such as self-reflection on existing knowledge and the ability to formulate new representations and new learning tasks that enable the learner to avoid stagnation and performance plateaus.

As reported in Michell *et al* (2018) NELL has been learning to read the web 24 hours/day since 2010, and at that time had acquired a knowledge base with over 80 million confidence weighted beliefs (*e.g*., servedWith(tea, biscuits).90 confidence). NELL has also learned millions of features and parameters that enable it to read these beliefs from the web. Additionally, it has learned to reason over these beliefs to infer (we might say using CSKR) new beliefs, and is able to extend its ontology by synthesizing new relational predicates. NELL learns to acquire knowledge in a variety of ways. It learns free-form text patterns for extracting this knowledge from sentences on a large scale corpus of web sites and it learns probabilistic rules that enable it to infer new instances of relations from other relation instances that it has already learned[viii]. As an example, NELL might learn a number of facts from a sentence defining "icefield", such as:

"a mass of glacier ice; similar to an ice cap, and usually smaller and lacking a dome-like shape; somewhat controlled by terrain."

In the context of this sentence and this new "background knowledge" extracted it might then extract supporting facts/particulars from following sentences:

"Kalstenius Icefield, located on Ellesmere Island, Canada, shows vast stretches of ice. The icefield produces multiple outlet glaciers that flow into a larger valley glacier."

Also of importance is that not only the textual situation is used to inter-relate extracted facts, but the physical location (*e.g*., Ellesmere Island) and any temporal situations expressed in these statements is used as context.[ix] NELL remains an example of how NLP and ML approaches can be used to build CSK and domain knowledge, but source context as well as ontology context needs to be taken into account to move forward. But NELL while it has extensive knowledge, it has relatively shallow semantic representations and thus suffers from ambiguities and inconsistencies (Gunning, 2018). And compared to handcrafted information such parts of extracted information are inconsistent with other parts and much noisier. Further, it is challenging to capture relevant situational context which include potentially important relations to other concepts - much of what is needed may be implicit and inferred and is currently only available in unstructured and un-annotated forms such as free text. And often training inputs to the model are highly engineered features that are complex or difficult to understand, meaning the resulting model learned will be hard to decompose for understanding use as input to explanation.

But progress on this problem comes from advanced ML applications where prior knowledge (background knowledge) may be used to judge the relevant facts of an extract, which makes this a bit of a bootstrapping situation.

Despite the remaining problems it seems reasonable that the role of existing and emerging Semantic Web technologies and associated ontologies is central to making CSKR population viable and that some extraction processes using a core of CSKR may be a useful way of proceeding.

CSKR helps Understanding and Thus Performance as well as Explanation in Contemporary ML Applications

As we have seen, the context that is important for discussing contemporary approaches to CSKR and explanations is that AI systems increasingly use advanced techniques such as deep learning (DL). These may in turn require additional techniques to make them more understandable to humans and system designers as well as trusted. For a different reason the current excited emphasis on explanation grows in part out of a feature failure of Deep Learning (DL) solutions - without additional effort they are opaque, at least in the sense that the models learned are not transparent to users or engineers. Despite this, contemporary deep neural networks (DNNs) have seemingly achieved near-human accuracy levels in various types of classification and prediction tasks including image and object recognition, text, speech, video data and behavior monitoring. These are all considered “low-level” tasks and advanced operations like planning or focused attention are not involved. Like simple rule-based explanations before them, raw DL systems do not natively handle desired aspects of explanations. Post-hoc explainability may be added to make them seem responsive. More recently, researchers, such as part of DARPA’s XAI program, as described by (Srihari, 2020) in this Issue, aim to create a suite of rich ML techniques that:

- produce more explainable models while maintaining a high level of learning performance and,

- enable humans to understand, appropriately trust, and effectively manage the emerging generation of AI "associates" that can be used in "high-level" domains such as healthcare, criminal justice system, and finance (Goodfellow 2016).

A notional architecture of a modern, hybrid intelligent system is shown in Figure 2. Here knowledge and reasoning are divided into several types which produce not only better problem solving abilities but explanations interpretable to a range of audience types. In order to achieve this a range of knowledge sources is involved as well as ML applications to further enrich the acquisition process.

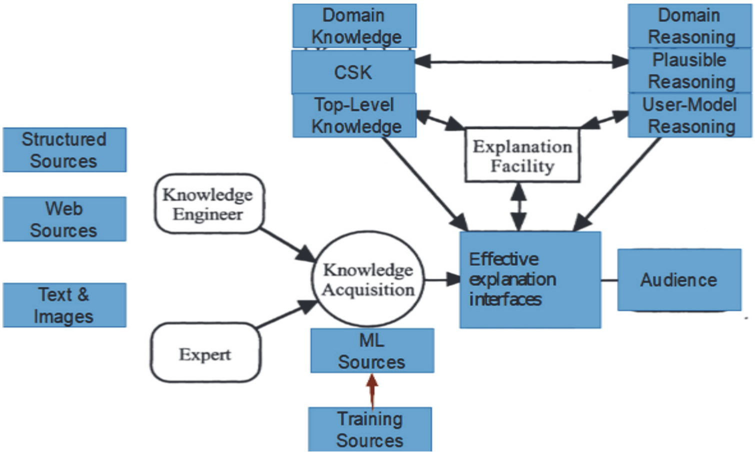


Figure 2: Architecture of a Hybrid Intelligent System

As an example, until recently the networks developed by ML for even simple vision detection approaches were treated mostly as black-box function approximators, in which a given input is mapped to some classification output such as the task of labeling images or translating text, as discussed in tracks of the 2019 Ontology Summit (Baclawski, 2018). So while ML and DL applications are now in wide use for common tasks such as advanced navigation with some sort of explanations to users, they are not naturally conducive to the generation of explanation structures. Because of complexity model simplification, say creating a decision tree, and/or feature relevance techniques which gauges the influence, relevance, or sensitivity each feature has in the prediction output by the model to be explained.

Supplications are often needed as a basis for explanations (Arrieta *et al,* 2020). Thus while non-technical, but valid and commonsense fashion are increasingly desired, they do not come without additional effort. Yet as Gunning summarizes:

> Machine reasoning is narrow and highly specialized. Developers must carefully train or program systems for every situation. General commonsense reasoning remains elusive. The absence of common sense prevents intelligent systems from understanding their world, behaving reasonably in unforeseen situations, communicating naturally with people, and learning from new experiences. Its absence is perhaps the most significant barrier between the narrowly focused AI applications we have today and the more general, human-like AI systems we would like to build in the future. (Gunning, 2018)

In a sense this is a return to an early desire to have smart applications knowledgeable about common phenomena and coincidentally ones capable of providing satisfying, interpretable explanations, but now positioned to take advantage of AI advances using DL. The path is necessary even though we still have not solved all the challenges of CSKR. Considering the range of application anticipated the goal of a reasonably competent CSKR system should include the ability to reason about explanations ("that makes sense") taking into account things like predictions, generalization, metaphors and abstractions, examples, as well as the goodness of plans, and diagnosis.

There is an obvious trust benefit if semi- or fully-automatic explanations can be provided as part of decision support systems. This seems like a natural extension of some long used and understood techniques such as logical proofs. Benefits can easily be seen if rich and deep deductions could be supported in areas regarding policies and legal issues, but also as part of automated education and training, such as e-learning. But there remains an inherent tension between ML performance (for example, predictive accuracy) as well as ideas of fairness and explainability. Often the highest-performing methods (for example, DL) are the least explainable, and the most explainable (for example, decision trees) are the least accurate and do not take into account the needs of the user.

Respective of formalisms and computational methods, an important criteria driving development is to ask “do these explanations make something clear?” DL systems are opaque and do not fully handle desired aspects of explanation to make them humanly comprehensible, which is the ability, in this case of ML algorithm to represent what is learned in a human understandable fashion. As noted, technical views may provide an answer to “how” the explanation was arrived at in steps and which rules or features were involved, but not the justifying and clarifying “why” of a satisfactory explanation. If, for example, a tree or hierarchical structure is involved in an explanation process we might get more of a “why” understanding with the possibility of drilling down and browsing a decision tree, having a focal point of attention on critical information or having the option of displaying a graphic representation that is human understandable. An example would be if a vehicle controller AI system for driving, based on visual sensing of objects could provide commonsense explanations (Persaud *et al*, 2017). Using internal commands a system may describe itself spatially as “moving forward”, while a human description is the more functional and just one of “driving down the street.” For explaining a lane change the system says, “because there are no other cars in my lane” while the human explanation is informative in another way “because the lane is clear.” These are similar but “clear” is a more comprehensive idea of a situation which might include construction, tree litter *etc*. (Tandon *et al*. 2018). A comprehendible explanation includes coherent pieces of information, more or less directly interpretable in natural language, and might relate quantitative (“no cars” and qualitative concepts (“near my lane”) in an integrated fashion.

It is important to note that under the influence of modern ML and Deep Learning (DL) models both CSKR and smart system explanations have recently been developing alongside these efforts and provide mutual support by co-developing deep explanations. These amount to modified or hybrid DL techniques that learn more explainable and CSKR features or representations or that feed into explanation generation facilities.

An area where we might see this developed is in the ability of DL-based applications to describe images (Geman, *et al*. 2015). This might be considered as one element of a visual Turing Test-like application and involve question- answering based on real-world images, such as detecting and localizing instances of objects and relationships among the objects in a scene. Some commonsense-making involved localizing questions[x] posed might include the following:

- What is Person1 carrying?
- What is Person2 placing on the table?
- Is Person3 standing behind a desk?
- Is Vehicle 1 moving?

This spate of recent work, reflecting the ability of ML systems to learn and answer questions about visual information and even text, has led to more distinctions being made about CSKR in support of robustness of the many ML applications which are increasingly thought of as mature enough to use for some ordinary tasks. Visual recognition is one of these, and supporting research approaches generate image captions to train a second deep network that can in turn generate explanations without explicitly identifying the original network's semantic features. This work continues but Shah *et al* (2019) suggests that some current ML applications are not robust as simple alternative NL syntactic formulations that lead to different answers. For example, if a system is asked "What is in the basket" and "What is contained in the basket" (or "what can be seen inside the basket") we get very different answers. Humans understand these as having similar commonsense meanings, but ML systems may have learned something different. And we may not know what they have learned and thus any direct explanation may be unsatisfactory for a user.

An obvious problem is that DL using a combination of efficient learning algorithms working over huge parametric space by themselves, are complex black-boxes in nature (Castelvecchi, 2016). For example, in a large knowledge graph measurements like nearest neighbors cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model. So while these approaches allow powerful predictions, their raw outputs cannot easily be directly explained and post hoc efforts are sometimes used. Consider the capability people have to distinguish the visual modality expressing a simply observed property like color or what seems like some simple relation like part. These afford common-sense and practical implications like "shiny things imply smoothness and so less friction". Distinctions like "smoothness" can play a role in transfer of training to new areas. Research now reliably shows the value of transfer training/learning such as with NELL. Transfer is enabled by pre-training a neural network model working on a known task, say image recognition using stored images from a general source like ImageNet. The resulting trained neural network (with an implied "model") is aimed for use with new, but related and purpose-specific models. What makes transfer difficult is finding training data that can provide a base to transfer for multiple types of scenarios and new situations of interest. There remain problems of representativeness and the selection of the typical to some generalizations such "shiny surfaces are typically hard, but some are not". There is also the problem of perspective. Imagine that we have in an image, the moon in the sky and a squirrel under a tree. They may seem the same size, but we know from common experience that they are at different distances and thus only appear to have a similar size. This is not something learned by a regular NN application, but it would be good to acquire this type of CSKR to allow this understanding.

Summing up Findings, Directions and Future Work

It seems clear that both CSKR and explanation remain important topics as part of AI research and its surging branch of ML. Further they can be mutually supportive, although explanation may be the more active area of diverse work just now. A guiding idea is that a truly explainable model should not have such knowledge gaps that users are left to generate different interpretations depending on their background knowledge. Having a suitable store of CSK can help an intelligent system produce explanations including natural language forms combining CSKR and human-understandable features (Bennetot, 2019).

For future direction five areas are noted:

1. Challenges in developing an adequate base of CSK
2. Situational and contextual understanding
3. Deep learning and dynamic situations
4. Interactions with humans
5. The need for a common, enhanced ontology engineering practice.

Adequate Knowledge: Providing a suitable base of CSK remains a broad, deep, and some say a largely unbounded problem. It seems generally true that one master ontology will not suffice for either specific domains or CSK and that a range of ontologies will be needed for an adequate CSK. Single ontologies are not likely to be suitable as work expands and more contexts are encountered. This will require multiple ontologies and/or a range of MTs as in Cyc. Big Knowledge, with its heterogeneity and depth complexity, may be as much of a problem as Big Data especially if we are leveraging heterogeneous, noisy, and conflicting data to create CSKR and explanations (Pauleen and Wang 2017). Various approaches do exist for different forms of CSKR, but the integration of these as well as ontologies with different content is still challenging. Linked data have a simplified view of KBs as a set of linked sentence-like assertions. However, integration of these requires some degree of background knowledge to understand the underlying assertions expressed in natural language labels. It is hard to imagine that major integration challenges from various forms with varying degrees of formality can be avoided. The ontology experience is that as a model of the real world we need to select some part of reality based on interest and conceptualization of that interest. Differences of selection and interpretation are impossible to avoid and it can be expected that different external factors will generate different contexts for any intelligent agent doing the selections and interpretation needed as part of a domain explanation.

The work such as Yi and Michael Gunginer, 2018 (Gunninger, 2018) suggests some coordinated set of ontologies that might be needed to support something as reasonable and focused as a Physical Embodied Turing Test. These include several aspects of intelligence, such as perception, reasoning, and action. Grunninger's suite (Gruninger, 2019), called PRAxIS (Perception, Reasoning, and Action across Intelligent Systems) with the following components:

- Solid Physical Objects (SoPhOs)
- Occupy (Location - Occupation is a relation between a physical body and a spatial region)
- Process Specification Language (PSL)
- Processes for Solid Physical Objects (ProSPerO)
- Ontologies for Video (OVid)
- Foundational Ontologies for Units of measure (FOUnt)

It is worth mentioning that ontologies like SoPhOs might emulate the intuitive physics of child cognition for objects while an "Occupy" concept provides notions of location and place used for spatial navigation. While this remains an early effort it does illustrate some of the diverse types of CSKR that need to be formalized. Yet there exists a range of strategies that could be employed to make progress on both the CSKR challenge and in its use to enhance explanations. In the sub-sections below, some of the remaining explanation and CSKR issues are further illustrated arising from some old problems that may affect more on the relatively newer challenges raised by ML and DL approaches.

Situational and contextual understanding: More complex tasks will involve greater situational understanding[xi]. These include situations where important things are unseen, but implied in a picture as part of the larger or assumed context such as exist in environmental or ecological settings with many dependencies. An example offered by Niket Tandon (Tandon *et al*., 2018) involves the implication of a directional arrow in a diagram of food web which intentionally communicates "consumes" to a human (a frog consumes bugs). The problem for modern learning oriented systems is that they are unlikely to have arrows used visually this way enough to generalize to a "consumes" meaning. To a human this is background knowledge.

Alas, it remains a hard problem to engineer all such knowledge or acquire it in an automated fashion. Indeed, since their inception, both explanatory systems and commonsense R & D have proven to involve implied, hard problems addressed by natural biological evolutions over a long period of time: such as the ideas of effective communication, consensus reality, background knowledge, notions of causality, and rationales. These allow the handling of things like focus and scale that is a known problem in visual identification. In a lake scene with a duck a ML vision system may see water features like dark spots as objects. In this case there seems a need for a model of the situation and for what is the focus of attention – a duck object. Some use of commonsense as part of model-based explanations might help during model debugging and decision making to correct apparently unreasonable predictions.

Such problems seem simple only because these are ubiquitous in everyday thinking, speaking, and perceiving as part of ordinary human interaction with the world. And this knowledge and reasoning seems easily captured because it is commonly available to the overwhelming majority of people, and manifest in human children's behavior by the age of three or five.

Deep Learning and Dynamic Situations Generally, the current state of the art for ML suggests that deep learning can provide some explanations of what they identify in simple visual datasets such as Visual Query Answering (VQA) and CLEVR. They can answer questions like "What is the man riding on?" in response to an image such as the one in Figure 3.

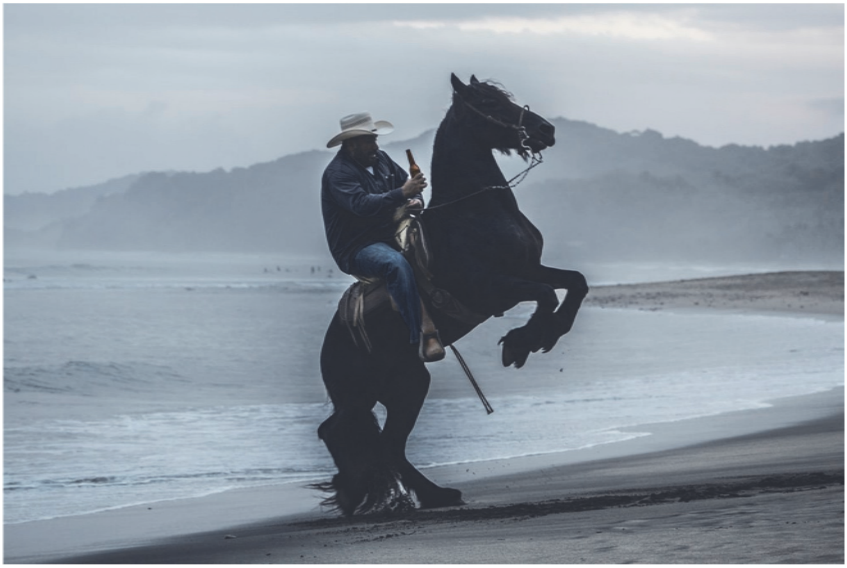


Figure 3: Example of image for ML processing

Whereas, commonsense knowledge is more important when the visual compositions are more dynamic and involve multiple objects and agents typical of say a cattle roundup. For dynamic and other situations further advanced intelligent system evolution needs to consider other features that may be supportive. This is true even in leaning-oriented systems like NELL which extract information from sentences. Because of things like contextual relations there remain many problems with un-sophisticated textual understanding. Examples are the implications and scope of negations and what is entailed.[xii]

Beyond negations there may be many situations one needs to understand – for example, "what exactly is happening in this ecological view?" This is challenging because a naive, start from scratch computational system, has to track everything involved in a situation or event. This may involve a long series of events with many objects and agents as in an ecological example or a food chain. Previously discussed situational complexity is also evident in visualizing a routine procedural play in basketball even as simple as a completed or missed dunk (Mishra *et al.*, 2019). Images of a dunk attempt can be described by three NL sentences: "He charges forward. And made a great leap. He made a basket." These sentences may be understood in terms of some underlying state-action-changes with a sequence of actions such as running and jumping, but there are also implied states as follows:

- The ball is in his hands. (not actually said, but seen and important for the play)
- The player is in the air. (implied by the leap)
- The ball is in the hoop. (technically how a basket is made)

We can represent the location of things in the three sentences above like this:

- Location (ball) = player's hand
- Location (player) = air
- Location (ball) = hoop (after Tanden, 2019)

These all fit into a coherent action with the context of a basketball script that we know, and thus humans can focus on the fact that the location of the ball at the end of the jump is a key result. CSKR about bodily capabilities apply here (Can I reach that hoop by jumping?) On other hand, as shown by Tandon *et al.* (2018), it is expensive to develop a large enough training set for such CSKR of activities, and the resulting state-action-change models have so many possible inferred candidate structures (*e.g.*, is the ball still in his hand? Maybe it was destroyed) so that common events can evoke an NP-complete problem. Without sufficient data (remember it is costly to construct), the model can produce what one would consider to be absurd, unrealistic choices based on commonsense experience such as the player being in the hoop.

A solution is to have a commonsense aware system that constrains the search for plausible event sequences. This is possible with the design and application of a handful of universally applicable rules. And some constraining ruling can be derived from existing ontologies. For example these constraints seem reasonable based on commonsense:

1. An entity must exist before it can be moved or destroyed. (certainly not likely in basketball)
2. An entity cannot be created if it already exists.
3. A tennis player is located at a tennis court.

In the work discussed by Niket Tanden (2018) these constraints were directly derivable from the Suggested Upper Merged Ontology (SUMO) rules such as: MakingFn, DestructionFn, MotionFn. This provides preliminary evidence that ontologies, even early ones such as SUMO, could be good guides for producing a handful of generic hard constraints in new domains.

One might ask, “How much help do these constraints provide?” The answer is that CSK-based search improves precision by nearly 30% over State-Of-The-Art DL efforts which include Recurrent Entity Networks (EntNet) , Query Reduction Networks (QRN) , and ProGlobal (Tanden *et al*, 2018).

Humans in the Loop While we do seem close to AI systems that will do common tasks such as driving a car or give advice on common tasks like eating it remains a challenge that such everyday tasks exhibit robust CSK and reasonable reasoning in order to be trusted. Monitoring the reasonableness and safety of automated actions, like driving in dynamic or even novel situations, illustrate a rapidly approaching but still challenging commonsense service capability. As intelligent agents become more autonomous, sophisticated, and prevalent, it becomes increasingly important that their knowledge become more complete and that humans be able to interact with them effectively to answer such questions as “why did you (my self-driving vehicle) take an unfamiliar turn?”

We need humans in the loop and allow dynamic interactions with intelligent agents. It is widely agreed that we need to enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners (Arrieta, 2020). Defining a successful application and its explanations remains relative to its audiences and their understanding. This is a bit of a psychological task so we can’t expect system designers and engineers to solve this without help (Mueller *et al*, 2019). But engineers can understand that human interactions and reactions to poor explanations can help to detect, and thus, correct things like bias in the training dataset or in system reasoning.

Current AI systems are good at recognizing objects, but can't explain what they see in ways understandable to and somewhat explainable by laymen. Nor can systems read a textbook and understand the questions in the back of the book which leads researchers to conclude they are devoid of common sense. We agree, as DARPA's Machine Common Sense (MCS) proposal put it that the lack of a common sense is "perhaps the most significant barrier" between the focus of AI applications today (such as previously discussed), and the human-like systems we dream of. And at least one of the areas that such an ability would play is with useful explanations. It may also be true, as NELL researchers argue, that we will never produce true NL understanding systems, until we have systems that react to arbitrary sentences with "I knew that, or didn't know and accept or disagree because X".

Better Methods for Engineering CSKR and Explanation: It is also worth noting that as explanation and CSKR research converge there is a need to develop a common, enhanced ontology engineering practice. As we arrive at a more focused understanding of CSKR there will be a need for this convergence to be incorporated into common ontological engineering practices. For efforts like CSK base building this should include guidance and best practices for the extraction of rules from extant, quality ontologies. A particular task is evaluating the quality of knowledge, both CSK and domain knowledge extracted from text. If knowledge is extracted from text and online information building of CSK will require methods to clean, refine and organize them. It is not as simple as saying that a system provides an exact match of words to what a human might produce given the many ways that meaning may be expressed. And it is costly to test system generated explanations or even captions against human ones due to the human cost.

One interesting research approach is to train a system to distinguish human and ML/DL system generated captions (for images *etc*.). After training one can use the resulting learned distinguished systems to critique the quality of the ML/DL generated labels.

In some cases, and increasingly so, a variety of CSK/information extracted is aligned (*e.g*. some information converges from different sources) by means of an extant (hopefully of high quality) ontology and perhaps several. This means that some aspect of the knowledge in the ontologies provides an interpretive or validating activity for the structuring involved in building artifacts like KGs. Knowledge graph gaps can also be filled in by internal processes looking for such things as consistency with common ideas as well as from external processes which adds information from human and/or automated sources. KG building efforts, which started employing sources like Freebase's data as a "gold standard" to evaluate data in DBpedia which in turn is used to populate a KG, are moving on to augmentation from text sources. In this light we can again note that a key requirement for validated table quality of knowledge involves the ability to trace back from a KB to the original source documents (such as LinkedData) and if filled in, from other sources such as humans to make it understandable or trustworthy. It is useful to note that this process of building such popular artifacts as KGs clearly shows that they are not equivalent in quality to supporting ontologies. In general there is some confusion in equating the quality of extracted information from text, KGs, KBs, the inherent knowledge in DL systems and ontologies.

But all such efforts are very probably going to rely on the assistance of new as yet undeveloped tools. In light of this future work we will need to refine a suite of tools and technologies to make the lifecycle of commonsense KBes easier and faster to build.

A successfully engineered intelligent system would be more of an "Associate Systems" with which users dialog with and over time get satisfactory answers because they include a capability to adaptively learn user knowledge and goals and are accountable for doing so over time. This is, of course, commonly true for human associates. The idea here is to mirror the user's mental model including some idea of commonsense, which becomes one of the main building block of intelligible human–machine interactions. Such focused, good, fair explanations may use natural language understanding to be part of a conversational dialogue human-computer interaction (HCI) in which the system uses previous knowledge of user (audience) knowledge and goals to discuss output explanations.

In such associate systems an issue will be the focus of attention. As part of common experience focus is an important element of explanations and commonsense assumptions and presumptions in a knowledge store play an important role in focus point. Indeed the ability to focus on relevant points may be part of the way a system competence is judged. But good focus has many potential dimensions and can involve judging and evaluating technical factors such as ethicality, fairness, and, where relevant, legality along with various roles such as relational, processual role, and social roles. These will all be important aspects of advanced AI applications. An example of this is that the role of legal advice is different in the context of a banking activity as opposed to lying under oath.

Acknowledgements

I thank Todd Schneider for a friendly review and suggestions.

References

1. Arrieta, Alejandro Barredo, *et al*. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* **58** (2020): 82-115.
2. Baclawski, K., Bennett, M., Berg-Cross, G., Casanave, C., Fritzsche, D., Luciano, J. & Sriram, R. D. (2018). Ontology Summit 2018 Communiqué: Contexts in context. *Applied Ontology*, **13**(3), 181-200.
3. A. Bennetot, J.-L. Laurent, R. Chatila, N. D´ıaz-Rodr´ıguez, Towards explainable neural-symbolic visual reasoning, in: NeSy Workshop IJCAI 2019, Macau, China, 2019.
4. Bundy, Alan. "A critique of proof planning." *Computational Logic: Logic Programming and Beyond*. Springer, Berlin, Heidelberg, 2002. 160-177.
5. Castelvecchi, D. Can we open the black box of AI? *Nature News* 538 (7623) (2016) 20.
6. Chakraborty, Supriyo, et al. "Interpretability of deep learning models: a survey of results." *2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*. IEEE, 2017.
7. Cohen, Robin, et al. "Trusted AI and the contribution of trust modeling in multiagent systems." *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2019.
8. Clancey, William J. "The epistemology of a rule-based expert system—a framework for explanation." *Artificial intelligence* 20.3 (1983): 215-251.
9. Darlington, Keith. "Aspects of intelligent systems explanation." *Universal Journal of Control and Automation* 1.2 (2013): 40-51.

10. Davis, Ernest, and Gary Marcus. "Commonsense reasoning and commonsense knowledge in artificial intelligence." *Communications of the ACM* 58.9 (2015): 92-103.

11. Fox, Maria, Derek Long, and Daniele Magazzeni. "Explainable planning." arXiv preprint arXiv:1709.10256 (2017).

12. Geman, Donald, *et al*. "Visual turing test for computer vision systems." *Proceedings of the National Academy of Sciences* 112.12 (2015): 3618-3623.

13. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016, http://www.deeplearningbook.org

14. Grosof, Benjamin, *et al*. "Automated decision support for financial regulatory/policy compliance, using textual rulelog." *Financial Times* (2014).

15. Gunning, David. "Machine common sense concept paper." *arXiv preprint arXiv:1810.07528* (2018).

16. Gruninger, Michael, Ontologies for the Physical Turing, Ontology Summit 2019 January, 2019 https://s3.amazonaws.com/ontologforum/OntologySummit2019/Commonsense/summit-physical-turing.pdf

17. Havasi, Catherine. "Reflections on Structured Common Sense in an Era of Machine Learning." *Proceedings of the 10th International Conference on Knowledge Capture*. 2019.

18. Holland, John H. "Escaping brittleness." *Proceedings Second International Workshop on Machine Learning*. 1983.

19. Kang, Dongyeop, *et al*. "Bridging Knowledge Gaps in Neural Entailment via Symbolic Models." *arXiv preprint arXiv:1808.09333* (2018).

20. Langley, Pat, *et al*. "Explainable Agency for Intelligent Autonomous Systems." AAAI. 2017.

21. Langlotz, C., and Shortliffe, E. Adapting a Consultation System to Critique User Plans. In Coombs, M. (Editor). Developments in Expert Systems. Academic Press, London. 1984.

22. Letham, B.; Rudin, C.; McCormick, T. H.; and Madigan, D. 2015. Interpretable Classifiers Using Rules and Bayesian Analysis: Building a Better Stroke Prediction Model. Annals of Applied Statistics

23. Lenat, Douglas B., Mayank Prakash, and Mary Shepherd. "CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks." *AI magazine* 6.4 (1985): 65-65.

24. Lenat, Douglas B., and Ramanathan V. Guha. *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc., 1989.

25. Lifschitz, Vladimir. "Formalizing Common Sense: Papers by John McCarthy." Ablex, Norwood, NJ (1990).

26. Lipton, Z.C.: The mythos of model interpretability. Workshop on Human Interpretability in Machine Learning (2016)

27. McCarthy, John. *Programs with common sense*. RLE and MIT computation center, 1960

28. McGuinness, Deborah L., and Paulo Pinheiro Da Silva. "Infrastructure for web explanations." *International Semantic Web Conference*. Springer, Berlin, Heidelberg, 2003.

29. Melis, Erica. "AI-techniques in proof planning." *The planner* 1.2 (1998): 3.

30. Miller, Tim. "Explanation in artificial intelligence: insights from the social sciences." arXiv preprint arXiv:1706.07269 (2017).

31. Minsky, Marvin. "Commonsense-based interfaces." *Communications of the ACM* 43.8 (2000): 66-73.

32. Minsky, Marvin L., Push Singh, and Aaron Sloman. "The St. Thomas common sense symposium: designing architectures for human-level intelligence." *Ai Magazine* 25.2 (2004): 113-113.

33. Mishra, Bhavana Dalvi, *et al*. "Everything Happens for a Reason: Discovering the Purpose of Actions in Procedural Text." *arXiv preprint arXiv:1909.04745* (2019).

34. Mitchell, Tom, *et al*. "Never-ending learning." *Communications of the ACM* 61.5 (2018): 103-115.

35. Moreau, Luc. "The foundations for provenance on the web." *Foundations and Trends® in Web Science* 2.2–3 (2010): 99-241.

36. Mueller, Shane T., *et al*. "Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI." *arXiv preprint arXiv:1902.01876* (2019).

37. Ontology Summit, https://ontologforum.org/index.php/OntologySummit2019 2019.

38. Ortiz Jr, Charles L. "Why we need a physically embodied Turing test and what it might look like." *AI magazine* 37.1 (2016): 55-62.

39. Pauleen, David J., and William YC Wang. "Does big data mean big knowledge? KM perspectives on big data and analytics." *Journal of Knowledge Management* (2017).

40. Panton, Kathy, *et al*. "Common sense reasoning–from Cyc to intelligent assistant." *Ambient Intelligence in Everyday Life*. Springer, Berlin, Heidelberg, 2006. 1-31.

41. Persaud, Priya, Aparna S. Varde, and Stefan Robila. "Enhancing autonomous vehicles with commonsense: Smart mobility in smart cities." *2017 IEEE 29th*

International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2017.

42. Singh, Push. "The open mind common sense project." *KurzweilAI. net* (2002a).
43. Singh, Push. "The public acquisition of commonsense knowledge." *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*. 2002b.
44. Shah, Meet, *et al*. "Cycle-consistency for robust visual question answering." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
45. Suchman, L. A. (1987). Plans and situated actions: The problem of human-machine communication. Cambridge university press.
46. Swartout, William R., and Stephen W. Smoliar. "Explanation: A source of guidance for knowledge representation." *Knowledge Representation and Organization in Machine Learning*. Springer, Berlin, Heidelberg, 1989. 1-16.
47. Tandon, Niket, Aparna S. Varde, and Gerard de Melo. "Commonsense knowledge in machine intelligence." *ACM SIGMOD Record* 46.4 (2018): 49-52.
48. Tandon, Niket, *et al*. "Reasoning about actions and state changes by injecting commonsense knowledge." *arXiv preprint arXiv:1808.10012* (2018).
49. Tandom, Niket Commonsense for Deep Learning, presented at Ontology Summit, 2019 http://people.mpi-inf.mpg.de/~ntandon/presentations/ontology-summit-2019/ontology-summit2019-niket-tandon.pdf
50. Von Rueden, Laura, *et al*. "Informed machine learning–towards a taxonomy of explicit integration of knowledge into machine learning." *Learning* 18 (2019): 19-20.
51. Walton, Douglas. *Abductive reasoning*. University of Alabama Press, 2014.
52. Yi, R. U., and Michael GR UNINGER. "What's the Damage? Abnormality in Solid Physical Objects." (2018).
53. Yuan, Changhe, Heejin Lim, and Tsai-Ching Lu. "Most Relevant Explanation in Bayesian Networks." *Journal Of Artificial Intelligence Research* (2011).

[i] The idea of intelligent systems covers a broad range of software technologies from simple heuristic and rule-based systems emulating human expertise with symbolic processing, to more recent neural network and machine learning technologies..

[ii] In "Programs with Common Sense" McCarthy (1960) described 3 tactical ways for early AI to proceed which includes common sense understanding and imitating the human central nervous system, which to a degree NN

systems do study human cognition or "understand the common sense world in which people achieve their goals."

iii In a narrow, logical and technical sense the "Gold Standard" concept of explanation is such a faithful, deductive proof done using a formal knowledge representation (KR).

iv These descended from J. McCarthy's tradition of treating contexts as formal objects over which one can quantify and express first-order properties.

v For example, "fact sentence F41 was drawn from document D73, at URL U84, in section 22, line 18." That kind of explanation is valuable and allows follow up.

vi Obviously a machine capability for a basic level of human-like commonsense would enable more effective communication and collaborate with their human partners.

vii As Andrej Karpathy put it, "I don't have to actually experience crashing my car into a wall a few hundred times before I slowly start avoiding to do so."

viii Reasoning is also applied for consistency checking and removing inconsistent axioms as in other knowledge graph (KG) generation efforts.

ix Knowledge reuse and transfer is an important issue in making such systems scalable.

x Broadly we might conceptualize this as a type of sensemaking in which an intelligent system that needs to analyze and interpret sensor or data input benefits from a CSKR service providing help it interpret and understand real world situations.

xi Some of these still unsolved contextual issues were discussed as part of the Ontology Summit 2018 on contexts (Baclawski *et al.*, 2018).

xii Kang *et al* (2018) showed the problems of what is concluded based on textual entailment with sentences from the Stanford Knowledge Language Inference set with sentences like "The dog did <u>not</u> eat all of the chickens."

BIO

Gary Berg-Cross is a cognitive psychologist (PhD, SUNY–Stony Brook) whose professional life included teaching and R&D in applied data & knowledge engineering, collaboration, and AI research. A board member of the Ontolog Forum he co-chaired the Research Data Alliance work-group on Data Foundations and Terminology. Major thrusts of his work include reusable knowledge, vocabularies, and semantic interoperability achieved through semantic analysis, formalization, capture in knowledge tools, and access through repositories.