Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence

Ernest Davis, Dept. of Computer Science, New York University

Gary Marcus, Dept. of Psychology, New York University

Abstract

Since the earliest days of artificial intelligence, it has been recognized that commonsense reasoning is one of the central challenges in the field. However, progress in this area has on the whole been frustratingly slow. In this review paper, we discuss why commonsense reasoning is needed to achieve human-level performance in tasks like natural language processing, vision, and robotics, why the problem is so difficult, and why progress has been slow. We also discuss four particular areas where substantial progress has been made, the techniques that have been attempted, and prospects for going forward.

Keywords: Commonsense reasoning, Artificial Intelligence, natural language processing, vision, robotics, knowledge base

1. Introduction

Artificial intelligence has seen great advances of many kinds recently, but there is one critical area where progress has been extremely slow: ordinary common sense.

Who is taller, Prince William or his baby son Prince George? Can you make a salad out of a polyester shirt? If you stick a pin into a carrot, does it make a hole in the carrot or in the pin? These types of questions may seem silly, but many intelligent tasks, such as understanding texts, computer vision, planning, and scientific reasoning require the same kinds of real-world knowledge and reasoning abilities. For instance, if you see a six-foot tall person holding a two-foot tall person in his arms, and you are told that they are father and son, you do not have to ask which is which. If you need to make a salad for dinner and are out of lettuce, you do not waste time considering improvising by taking a shirt of the closet and cutting it up. If you read the text, “I stuck a pin in a carrot; when I pulled the pin out, it had a hole,” you need not consider the possibility that “it” refers to the pin.

To take another example, consider what happens when we watch a movie, putting together information about the motivations of fictional characters we’ve met only moments before. Anyone who has seen the unforgettable horse’s head scene in *The Godfather* immediately realizes what’s going on. It’s not just that it’s unusual to see a severed horse head, it’s clear that Tom Hagen is sending Jack Woltz a message – if I can decapitate your horse, I can decapitate you; cooperate, or else. For now, such inferences lie far beyond anything in artificial intelligence.

Here, after arguing that commonsense reasoning is important in many AI tasks, from text understanding to computer vision, planning and reasoning (section 2), and discussing four specific problems where substantial progress has been made (section 3), we consider why the problem in its general form is so difficult and why progress has been so slow (section 4). We then survey various techniques that have been attempted (section 5) and conclude with some modest proposals for future research.

2. Common sense in intelligent tasks

2.1 Natural language processing

The importance of real-world knowledge for natural language processing, and in particular for disambiguation of all kinds, was discussed as early as 1960, by Bar-Hillel (1960), in the context of machine translation. Although some ambiguities can be resolved using simple rules that are comparatively easy to acquire, a substantial fraction can only be resolved using a rich understanding of the world. A well-known example, due to Terry Winograd (1972), is the pair of sentences "The city council refused the demonstrators a permit because they feared violence," vs."… because they advocated violence". To determine that "they" in the first sentence refers to the council if the verb is "feared" but refers to the demonstrators if the verb is "advocated" demands knowledge about the characteristic relations of city councils and demonstrators to violence; no purely linguistic clue suffices.[1]

Machine translation likewise often involves problems of ambiguity that can only be resolved by achieving an actual understanding of the text — and bringing real-world knowledge to bear. Google Translate often does a fine job of resolving ambiguities by using nearby words; for instance, in translating the two sentences "The electrician is working" and "The telephone is working" into German, it correctly translates "working" as meaning "laboring", in the first sentence and as meaning "functioning correctly" in the second, because in the corpus of texts that Google has seen, the German words for "electrician" and "laboring" are often found close together, as are the German words for "telephone" and "function correctly".[2] However if you give it the sentences "The electrician who came to fix the telephone is working," and "The telephone on the desk is working", interspersing several words between the critical element (e.g. between electrician and working), the translations of the longer sentences say that the electrician is functioning properly and that the telephone is laboring (Table 1). A statistical proxy for common sense that worked in the simple case fails in the more complex case.

[1] Such pairs of sentences are known as "Winograd schemas" after this example; a collection of many such examples can be found at http://cs.nyu.edu/faculty/davise/papers/WS.html (Levesque, Davis, & Morgenstern, 2012).

[2] Google Translate is a moving target; this particular example was carried out on 6/9/2015, but translations of individual sentences change rapidly —not always for the better on individual sentences. Indeed, the same query given minutes apart can give different results.. Changing the target language, or making seemingly inconsequential changes to the sentence, can also change how a given ambiguity is resolved, for no discernible reason. Our broader point here is not to dissect Google Translate per se, but to note that it is unrealistic to expect fully reliable disambiguation in the absence of a deep understanding of the text and relevant domain knowledge.

English original	Google translation
The electrician is working.	Der Electriker **arbeitet**.
The electrician that came to fix the telephone is working.	Der Elektriker, die auf das Telefon zu beheben kam **funktioniert**.
The telephone is working.	Das Telefon **funktioniert**.
The telephone on the desk is working.	Das Telefon auf dem Schreibtisch **arbeitet**.

Table 1: Lexical ambiguity and Google Translate.
We have highlighted the translation of the word "working". The German word "arbeitet" means "labors"; "funktioniert" means "functions correctly."

Almost without exception, current computer programs to carry out language tasks succeed to the extent that the tasks can be carried out purely in terms of manipulating individual words or short phrases, without attempting any deeper understanding; common sense is evaded, in order to focus on short-term results, but it's hard to see how human-level understanding can be achieved without greater attention to common sense.

Watson, the Jeopardy-playing program, is an exception to the above rule only to a small degree. As described in (Kalyanpur, 2012), commonsense knowledge and reasoning, particularly taxonomic reasoning, geographic reasoning, and temporal reasoning, played some role in Watson's operations but only a quite limited one, and they made only a small contribution to Watson's success. The key techniques in Watson are mostly of the same flavor as those used in programs like web search engines: there is a large collection of extremely sophisticated and highly tuned rules for matching words and phrases in the question with snippets of web documents such as Wikipedia; for reformulating the snippets as an answer in proper form; and for evaluating the quality of proposed possible answers. There is no evidence that Watson is anything like a general purpose solution to the common sense problem.

2.2 Computer Vision

Similar issues arise in computer vision. Consider the photograph of Julia Child's kitchen in Figure 1: Many of the objects that are small or partially seen, such as the metal bowls in the shelf on the left, the cold water knob for the faucet, the round metal knobs on the cabinets, the dishwasher, and the chairs at the table seen from the side, are only recognizable in context; the isolated image would be hard to identify. The top of the chair on the far side of the table is only identifiable because it matches the partial view of the chair on the near side of the table.

The viewer infers the existence of objects that are not in the image at all. There is a table under the yellow tablecloth. The scissors and other items hanging on the board in the back are presumably supported by pegs or hooks. There is presumably also a hot water knob for the faucet occluded by the dish rack. The viewer also infers how the objects can be used (sometimes called their "affordances") e.g., that the cabinets and shelves can be opened by pulling on the handles. (Cabinets, which rotate on joints, have the handle on one side; shelves, which pull out straight, have the handle in the center.)

Movies would prove even harder; few AI programs have even tried. The *Godfather* scene mentioned earlier is one example, but almost any movie contains dozens or hundreds of moments that cannot be understood simply by matching still images to memorized templates. Understanding a movie requires a viewer to make numerous inferences about the intentions of characters, the nature of physical objects, and so forth. In the current state of the art, it is not feasible even to attempt to build a program that will be able to do this reasoning; the most that can be done is to track characters and identify basic actions like standing up, sitting down, and opening a door (Bojanowski, et al., 2014)

Figure 1: Julia Child's kitchen.[3]

2.3 Robotic Manipulation

The need for commonsense reasoning in autonomous robots working in an uncontrolled environment is self-evident, most conspicuously in the need to have the robot react to unanticipated events appropriately. If a guest asks a waiter-robot for a glass of wine at a party, and the robot sees that the glass he has gotten is broken, or has a dead cockroach at the bottom, the robot should not simply pour the wine into the glass and serve it. If a cat runs in front of a house-cleaning robot, the robot should neither run it over nor sweep it up nor put it away on a shelf. These things seem obvious, but ensuring that a robot avoids mistakes of this kind is very challenging.

3. Successes in Automated Commonsense Reasoning

Substantial progress in automated commonsense reasoning has been made in four areas: reasoning about taxonomic categories, reasoning about time, reasoning about actions and change, and the sign calculus. In each of these areas there exists a well-understood theory that can account for some broad range of commonsense inferences.

[3] Photograph by Matthew Bisanz; reproduced under a Creative Commons License.
http://en.wikipedia.org/wiki/File:Julia_Child%27s_kitchen_by_Matthew_Bisanz.JPG

3.1 Taxonomic reasoning

A taxonomy is a collection of categories and individuals, and the relations between them. (Taxonomies are also known as semantic networks.)

For instance, figure 3 shows a taxonomy of a few categories of animals and individuals.

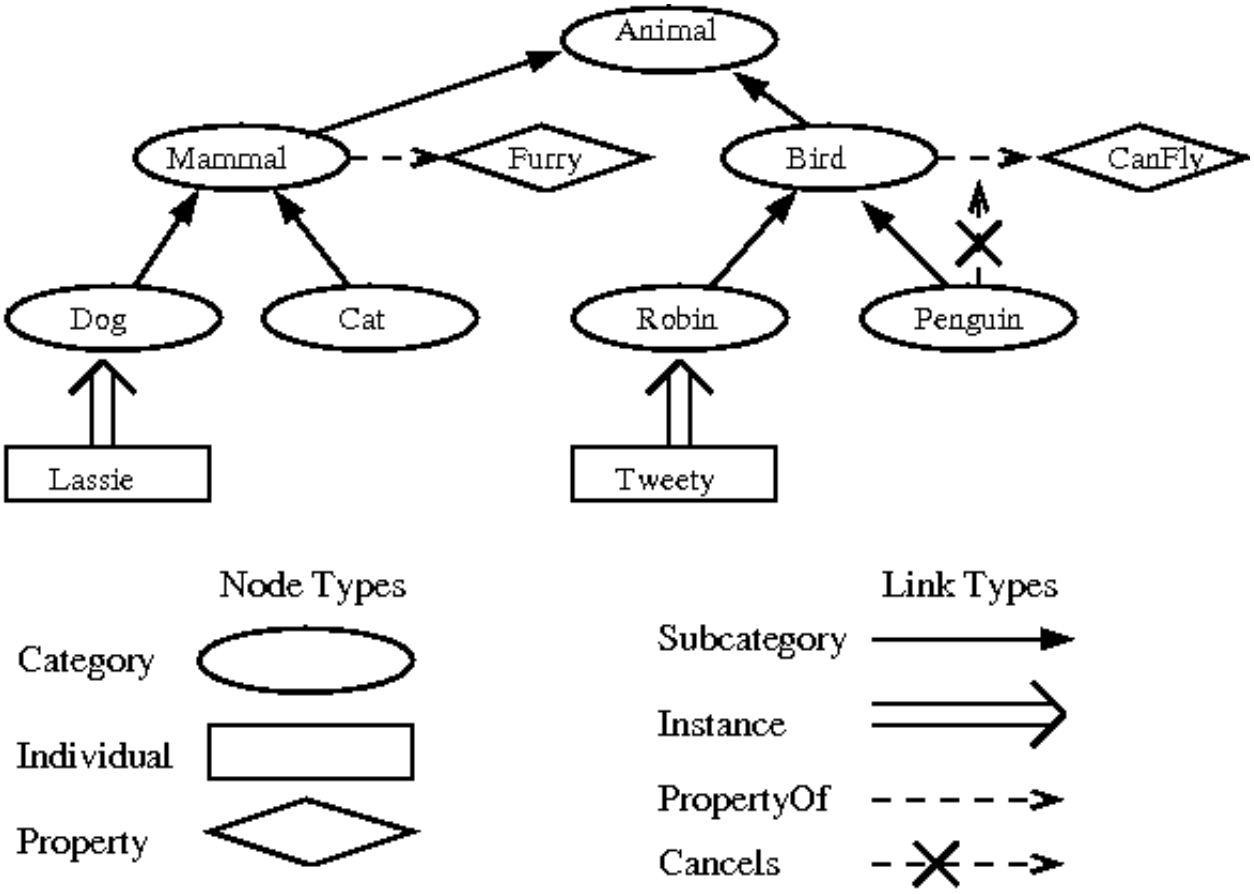


Figure 3: Taxonomy

There are three basic relations here:

- An individual is an instance of a category. For instance, the individual `Lassie` is an instance of the category `Dog`.
- One category is a subset of another. For instance `Dog` is a subset of `Mammal`.
- Two categories are disjoint. For instance `Dog` is disjoint from `Cat`.

Figure 3 does not indicate the disjointness relations.

Categories can also be tagged with properties. For instance, `Mammal` is tagged as `Furry`.

One form of inference in a taxonomy is transitivity. Since `Lassie` is an instance of `Dog` and `Dog` is a subset of `Mammal`, it follows that `Lassie` is an instance of `Mammal`. Another form of inference is inheritance. Since `Lassie` is an instance of `Dog` which is a subset of `Mammal` and `Mammal` is marked

with property Furry, it follows that Dog and Lassie have property Furry. A variant of this is default inheritance; a category can be marked with a characteristic but not universal property, and a subcategory or instance will inherit the property unless it is specifically cancelled. For instance Bird has the default property CanFly, which is inherited by Robin but not by Penguin.

The standard taxonomy of the animal kingdom is particularly simple in structure. The categories are generally sharply demarcated. The taxonomy is tree-structured, meaning that given any two categories, either they are disjoint or one is a subcategory of the other. Other taxonomies are less straightforward. For instance, in a semantic network for categories of people, the individual GalileoGalilei is simultaneously a Physicist, an Astronomer, a ProfessorOfMathematics, a WriterInItalian, a NativeOfPisa, a PersonChargedWithHeresy, and so on. These overlap, and it is not clear which of these are best viewed as taxonomic categories and which are better viewed as properties. In taxonomizing more abstract categories, choosing and delimiting categories becomes more problematic; for instance, in constructing a taxonomy for a theory of narrative, the membership, relations, and definitions of categories like Event, Action, Process, Development, and Incident are uncertain.

Simple taxonomic structures such as that illustrated above are often used in AI programs. For example, WordNet (Miller, 1995) is a widely used resource that includes a taxonomy whose elements are meanings of English words. As we will discuss in section 5.2, web mining systems that collect commonsense knowledge from web documents tend to be largely focused on taxonomic relations, and more successful in gathering taxonomic relations than in gathering other kinds of knowledge. Many specialized taxonomies have been developed in domains such as medicine (Pisanelli, 2004) and genomics (Gene Ontology Consortium, 2004). More broadly, the Semantic Web enterprise is largely aimed at developing architectures for large-scale taxonomies for web applications.

A number of sophisticated extensions of the basic inheritance architecture described above have also been developed. Perhaps the most powerful and widely used of these is description logic (Baader, Horrocks, & Sattler, 2008). Description logics provide tractable constructs for describing concepts and the relations between concepts, grounded in a well-defined logical formalism. They have been applied extensively in practice, most notably in the semantic web ontology language OWL.

3.2 Temporal Reasoning

Representing knowledge and automating reasoning about times, durations, and time intervals is a largely solved problem (Fisher, 2008). For instance, if one knows that Mozart was born earlier and died younger than Beethoven, one can infer that Mozart died earlier than Beethoven. If one knows that the Battle of Trenton occurred during the Revolutionary War, that the Battle of Gettysburg occurred during the Civil War and that the Revolutionary War was over before the Civil War started, then one can infer that the Battle of Trenton occurred before the Battle of Gettysburg. The inferences involved here in almost all cases reduce to solving systems of linear inequalities, usually small and of a very simple form.

Integrating such reasoning with specific applications, such as natural language interpretation, has been much more problematic. Natural language expressions for time are complex and their interpretation is context dependent. Temporal reasoning was used to some extent in the Watson Jeopardy-playing program to exclude answers that would be a mismatch in terms of date (Kalyanpur, 2012). However, many important temporal relations are not explicitly stated in texts, they are inferred; and the process of inference can be difficult. Basic tasks like assigning time-stamps to events in news stories cannot be currently done with any high degree of accuracy (Surdeanu, 2013).

3.3 Action and Change

Another area of commonsense reasoning that is well understood is the theory of action, events, and change. In particular, there are very well established representational and reasoning techniques for domains that satisfy the following constraints (Reiter, 2001):

- Events are atomic. That is, one event occurs at a time, and the reasoner need only consider the state of the world at the beginning and the end of the event, not the intermediate states while the event is in progress.
- Every change in the world is the result of an event.
- Events are deterministic; that is, the state of the world at the end of the event is fully determined by the state of the world at the beginning plus the specification of the event.
- Single actor. There is only a single actor, and the only events are either his actions or exogenous events in the external environment.
- Perfect knowledge. The entire relevant state of the world at the start, and all exogenous events are known or can be calculated

For domains that satisfy these constraints, the problem of representation and important forms of reasoning such as prediction and planning, are largely understood.

Moreover a great deal is known about extensions to these domains, including

- Continuous domains, where change is continuous.
- Simultaneous events.
- Probabilistic events, whose outcome depends partly on chance.
- Multiple agent domains, where agents may be cooperative, independent, or antagonistic.
- Imperfect knowledge domains, where actions can be carried out with the purpose of gathering information, and (in the multi-agent case) where cooperative agents must communicate information.
- Decision theory: Comparing different courses of action in terms of the expected utility.

The primary successful applications of these kinds of theories has been to high-level planning (Reiter, 2001), and to some extent to robotic planning e.g. (Ferrein, Fritz, & Lakemeyer, 2005).

The situation calculus uses a branching model of time, because it was primarily developed to characterize planning, in which one must consider alternative possible actions. However, it does not work well for narrative interpretation, since it treats events as atomic and requires that the order of events be known. For narrative interpretation, the event calculus (Mueller, 2006) is more suitable. The event calculus can express many of the temporal relations that arise in narratives; however, only limited success has been obtained so far in applying it in the interpretation of natural language texts. Moreover, since it uses a linear model of time, it is not suitable for planning.

Many important issues remain unsolved, however, such as the problem of integrating action descriptions at different levels of abstraction. The process of cooking dinner, for instance, may involve such actions as "Getting to know my significant other's parents," "Cooking dinner for four," "Cooking pasta primavera," "Chopping a zucchini," "Cutting once through the zucchini", and "With the right hand, grasping the knife by the handle, blade downward, and lowering it at about 1 foot per second through the center of the zucchini, while, with the left hand, grasping the zucchini and holding it against the cutting board." Reasoning about how these different kinds of actions interrelate — e.g. if you manage to slice off a finger while cutting the zucchini, your prospective parents-in-law may not be impressed — is substantially unsolved.

3.4 Qualitative reasoning

One type of commonsense reasoning that has been analyzed with particular success is known as qualitative reasoning. In its simplest form, qualitative reasoning is about the direction of change in interrelated quantities. If the price of an object goes up then (usually, other things being equal) the number sold will go down. If the temperature of gas in a closed container goes up, then the pressure will

go up. If an ecosystem contains foxes and rabbits and the number of foxes decreases, then the death rate of the rabbits will decrease (in the short term).

An early version of this theory was formulated by Johan de Kleer (1975) for analyzing an object moving on a roller coaster. Later more sophisticated forms were developed in parallel by de Kleer and Brown (1985) for analyzing electronic circuits; by Forbus (1985) for analyzing varieties of physical processes; and by Kuipers (1986) as a mathematical formalism.

This theory has been applied in many domains, from physics to engineering, biology, ecology, and engineering. It has also served as the basis for a number of practical programs, including text understanding (Kuehne (2004); analogical mapping and geometric reasoning (Lovett et al, 2009) failure analysis in automotive electrical systems (Price, Pugh, Wilson, & Snooke, 1995); and generating control strategies for printing (Fromherz, Bobrow, & de Kleer, 2003).

For problems within the scope of the representation, the reasoning mechanism works well. However, there are many problems in physical reasoning, particularly those involving substantial geometric reasoning, that cannot be represented in this way, and therefore lie outside the scope of this reasoning mechanism. For example, you want to be able to reason that a basketball will roll smoothly in any direction, whereas a football can roll smoothly if its long axis is horizontal but cannot roll smoothly end over end. This involves reasoning about the interactions of all three spatial dimensions together.

4. Challenges in automating commonsense reasoning

As of 2014, few commercial systems make any significant use of automated commonsense reasoning. Systems like Google Translate use statistical information culled from large data sets as a sort of distant proxy for commonsense knowledge, but beyond that sort of crude proxy, commonsense reasoning is largely absent. In large part, that is because nobody has yet come close to producing a satisfactory common-sense reasoner. There are five major obstacles.

First, many of the domains involved in commonsense reasoning are only partially understood or virtually untouched. We are far from a complete understanding of domains such as physical processes, knowledge and communication, plans and goals, and interpersonal interactions. In domains such as the commonsense understanding of biology, of social institutions, or of other aspects of folk psychology, little work of any kind has been done.

Second, situations that seem straightforward can turn out, on examination, to have considerable logical complexity. For instance, the horse's head scene in *The Godfather* the viewer understands that:

> Hagen foresaw, while he was planning the operation, that
> Woltz would realize that
> Hagen arranged for the placing of the head in Woltz's bed
> in order to make
> Woltz realize that
> Hagen could easily arrange to have him killed
> if he does not accede to Hagen's demands.

Thus we have a statement with embeddings of three mental states ("foresaw", "realize", "realize"), a teleological connection ("in order"), two hypotheticals ("could arrange" and "does not accede") and a highly complex temporal/causal structure.

Some aspects of these kinds of relations have been extensively studied and are well understood. However, there are many aspects of these relations where we do not know, even in principle, how they can be represented in a form usable by computers or how to characterize correct reasoning about them. For example, there are theories of knowledge that do a good job of representing what different players know about the deal in a poker game, and what each player knows about what the other players know,

because one can reasonably idealize all the players as being able to completely think through the situation. However, if you want to model a teacher thinking about what his students don't understand, and how they can be made to understand, then that is a much harder problem, and one for which we currently do not have a workable solution. Moreover, even when the problems of representation and inference have been solved in principle, the problem of carrying out reasoning efficiently remains.

Third, commonsense reasoning almost always involves plausible reasoning; that is, coming to conclusions that are reasonable given what is known, but not guaranteed to be correct. Plausible reasoning has been extensively studied for many years (Halpern, 2003), and many theories have been developed, including probabilistic reasoning (Pearl, 1988), belief revision (Peppas, 2008), and default reasoning or non-monotonic logic (Brewka, Nielmellä, & Truszczinski, 2008). However, overall we do not seem to be very close to a comprehensive solution. Plausible reasoning takes many different forms, including using unreliable data; using rules whose conclusions are likely but not certain; default assumptions; assuming that one's information is complete; reasoning from missing information; reasoning from similar cases; reasoning from typical cases; and others. How to do all these forms of reasoning acceptably well in all commonsense situations and how to integrate these different kinds of reasoning are very much unsolved problems.

Fourth, in many domains, a small number of examples are highly frequent, while there is a "long tail" of a vast number of highly infrequent examples. In natural language text, for example, some trigrams (e.g. "of the year") are very frequent, but many other possible trigrams, such as "moldy blueberry soda" or "gymnasts writing novels" are immediately understandable, yet vanishingly rare.[4] Long tail phenomena also appear in many other corpora, such as labeled sets of images (Russell, Torralba, Murphy, & Freeman, 2008).

The effect of long tail distributions on AI research can be pernicious. On the one hand, promising preliminary results for a given task can be gotten easily, because a comparatively small number of common categories include most of the instances. On the other hand, it is often very difficult to attain high quality results, because a significant fraction of the problems that arise correspond to very infrequent categories. The result is the pattern of progress often seen in AI: Rapid progress at the start of research up to a mediocre level, followed by slower and slower improvement. (Of course, for any given application, partial success may be acceptable or indeed extremely valuable; and high quality performance may be unnecessary.)

We conjecture that long tail phenomena are pervasive in commonsense reasoning, both in terms of the frequency with which a fact appears in knowledge sources (e.g. texts) and in terms of the frequency with which it is needed for a reasoning task. For instance, as discussed above, a robot waiter needs to realize that it should not serve a drink in a glass with a dead cockroach; but how often is that mentioned in any text, and how often will the robot need to know that fact?[5]

Fifth, in formulating knowledge it is often difficult to discern the proper level of abstraction. Recall the example of sticking a pin into a carrot and the reasoning that this action may well create a hole in the carrot, but not create a hole in the pin. Before it encounters this particular example, an automated reasoner presumably would not specifically know a fact specific to pins and carrots; at best it might know

[4] Google reports no instances of either of these quoted phrases as of June 9, 2015. These are not hard to find in natural text; for example, one recent book review that we examined contained at least eight trigrams (not containing proper nouns) with zero Google hits other than the article itself. A systematic study of n-gram distribution can be found in (Allison, Guthrie, & Guthrie, 2006).

[5] Presumably an intelligent robot would not necessarily know that specific fact in advance at all, but rather would infer it when necessary. However, accomplishing that involves describing the knowledge that supports the inference and building the powerful inference engine that carries out the inference.

a more general rule[6] or theory about creating holes by sticking sharp objects into other objects. The question is, how broadly should such rules should be formulated? Should such roles cover driving nails into wood, driving staples into paper, driving a spade into the ground, pushing your finger through a knitted mitten, or putting a pin into water (which creates a hole that is immediately filled in)? Or must there be individual rules for each domain? Nobody has yet presented a general solution to this problem.

A final reason for the slow progress in automating commonsense knowledge is both methodological (Davis, 1998) and sociological. Piecemeal commonsense knowledge (e.g. specific facts) is relatively easy to acquire, but often of little use, because of the long-tail phenomenon discussed above. Consequently, there may not be much value in being able to do a *little* commonsense reasoning. The payoff in a complete commonsense reasoner would be large, especially in a domain like robotics, but that payoff may only be realized once a large fraction of the project has been completed. By contrast, the natural incentives in software development favor projects where there are payoffs at every stage; projects that require huge initial investments are much less appealing.

5. Approaches and techniques

As with most areas of artificial intelligence, the study of commonsense reasoning is largely divided into knowledge-based approaches and approaches based on machine learning over large data corpora (almost always text corpora) with only limited interaction between the two kinds of approaches. There are also crowd-sourcing approaches, which attempt to construct a knowledge-base by somehow combining the collective knowledge and participation of many non-expert people. Knowledge-based approaches can in turn be divided into approaches based on mathematical logic or some other mathematical formalism; informal approaches, antipathetic to mathematical formalism, and sometimes based on theories from cognitive psychology; and large-scale approaches, which may be more or less mathematical or informal, but in any case are chiefly targeted at collecting a lot of knowledge. A particularly successful form of mathematically grounded commonsense reasoning is qualitative reasoning, described in section 3.4 above. We consider these in turn.

[6] Positing that the reasoner is not using rules at all, but instead is using an instance-based theory does not eliminate the problem. Rather, it changes the problem to the question of how to formulate the features to be used for comparison.

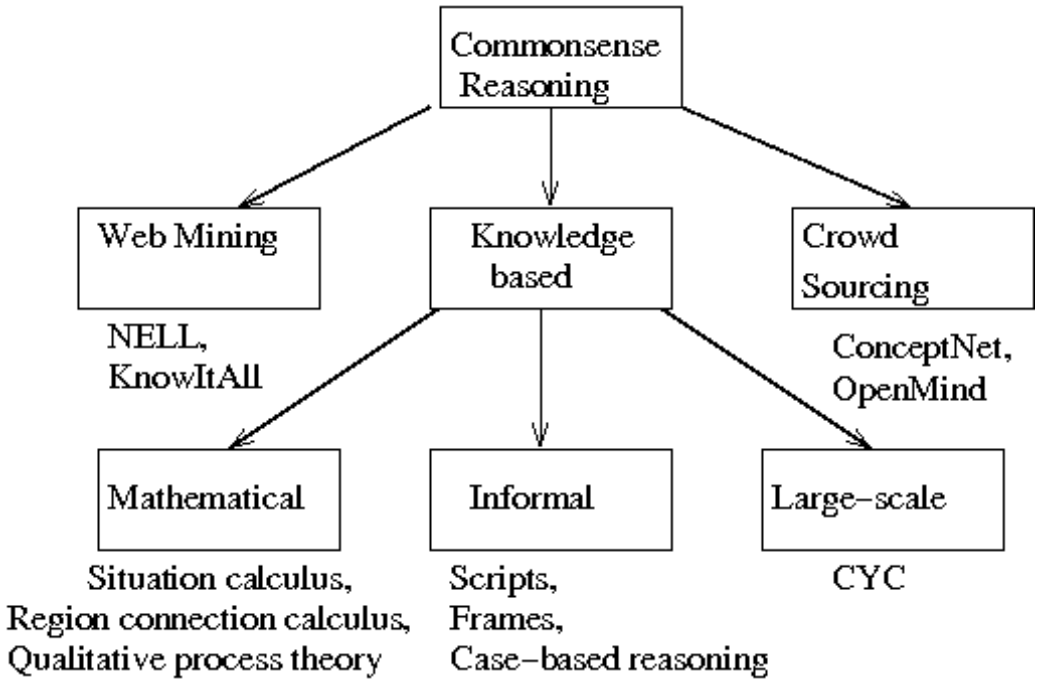


Figure 4: Taxonomy of approaches to commonsense reasoning

Research in commonsense reasoning addresses a number of different objectives:

- Reasoning architecture. The development of general-purpose data structures for encoding knowledge and algorithms and techniques for carrying out reasoning. (A closely related issue is the representation of the meaning of natural language sentences (Schubert, 2015).)
- Plausible inference; drawing provisional or uncertain conclusions.
- Range of reasoning modes. Incorporating a variety of different modes of inference, such as explanation, generalization, abstraction, analogy, and simulation.
- Painstaking analysis of fundamental domains. In doing commonsense reasoning, people are able to do complex reasoning about basic domains such as time, space, naïve physics, and naïve psychology. The knowledge that they are drawing on is largely unverbalized and the reasoning processes largely unavailable to introspection. An automated reasoner will have to have comparable abilities.
- Breadth. Attaining powerful commonsense reasoning will require a large body of knowledge.
- Independence of experts. Paying experts to hand-code a large knowledge base is slow and expensive. Assembling the knowledge base either automatically or by drawing on the knowledge of non-experts is much more efficient
- Applications. To be useful, the commonsense reasoner must serve the needs of applications and must interface with them smoothly.
- Cognitive modeling. Theories of commonsense automated reasoning accurately describe commonsense reasoning in people.

The different approaches to automating commonsense reasoning have often emphasized different objectives, as sketched in Table 2.

	Math based	Informal	Large -scale	Web mining	Crowd sourcing
Architecture	Substantial	Little	Substantial	Moderate	Little
Plausible reasoning	Substantial	Moderate	Substantial	Little	Little
Range of reasoning modes	Moderate	Substantial	Moderate	Little	Little
Painstaking fundamentals	Substantial	Little	Moderate	Little	Little
Breadth	Little	Moderate	Substantial	Substantial	Substantial
Independence of experts	Little	Little	Little	Substantial	Substantial
Concern with applications	Moderate	Substantial	Substantial	Moderate	Moderate
Cognitive modeling	Little	Substantial	Little	Little	Moderate

Table 2: Approaches and typical objectives

5.1 Knowledge-based approaches

In knowledge-based approaches, experts carefully analyze the characteristics of the inferences needed to do reasoning in a particular domain or for a particular task and the knowledge that those inferences depend on. They hand-craft representations that are adequate to express this knowledge and inference engines that are capable of carrying out the reasoning.

5.1.1 Mathematically grounded approaches

Of the four successes of commonsense reasoning enumerated in section 4, all but taxonomic reasoning largely derive from theories that are grounded in mathematics or mathematical logic. (Taxonomic representations are too ubiquitous to be associated with any single approach.) However, many directions of work within this approach are marked by a large body of theory and a disappointing paucity of practical applications or even of convincing potential applications. The work on qualitative spatial reasoning (Cohn & Renz, 2007) illustrates this tendency vividly. There has been active work in this area for more than twenty years, and more than a thousand research papers have been published, but very little of this connects to any commonsense reasoning problem that might ever arise. Similar gaps between theory and practice arise in other domains as well. The "Muddy Children" problem[7] (also known as the "Cheating Husbands" problem), a well-known brain teaser in the theory of knowledge, has been analyzed in a dozen different variants in a half-dozen different epistemic logics; but we do not know how to represent the complex interpersonal interactions between Hagen and Woltz in the horse's head scene, let alone how to automate reasoning about them.

[7] Alice, Bob, and Carol are playing together. Dad says to them, "At least one of you has mud on your forehead. Alice, is your forehead muddy?" Alice answers, "I don't know." Dad asks, "Bob, is your forehead muddy?" Bob answers, "I don't know.". Carol then says, "My forehead is muddy." Explain.

Unlike the other approaches to commonsense reasoning, much of the work in this approach is purely theoretical; the end result is a published paper rather than an implemented program. Theoretical work of this kind is evaluated, either in terms of meta-theorems (e.g. soundness, completeness, computational complexity), or in terms of interesting examples of commonsense inferences that the theory supports. These criteria are often technically demanding; however, their relation to the advancement of the state of the art is almost always indirect, and all too often nonexistent.

Overall the work is also limited in terms of the scope of domains and reasoning techniques that have been considered. Again and again, research in this paradigm has fixated on a small number of examples and forms of knowledge, and generated vast collections of papers dealing with these, leaving all other issues neglected.

5.1.2 Informal knowledge-based approaches

In the informal knowledge-based approach, theories of representation and reasoning are based substantially on intuition and anecdotal data, and to a significant but substantially lesser extent on results from empirical behavioral psychology.

The most important contribution of the informal approach has been the analysis of a broad class of types of inference. For instance, Minsky's frame paper (1975) discusses an important form of inference, in which a particular complex individual, such as a particular house, is matched against a known structure, such as the known characteristics of houses in general. Schank's theory of scripts (Schank & Abelson, 1977) addresses this in the important special case of structured collections of events. Likewise reasoning by analogy (Hofstadter & Sander, 2013) (Gentner & Forbus, 2011) and case-based reasoning (Kolodner, 1993) have been much more extensively studied in informal frameworks than in mathematically grounded frameworks.

It should be observed that in computer programming generally, informal approaches are very common. Many large and successful programs — text editors, operating systems shells, and so on — are not based on any overarching mathematical or statistical model; they are written ad hoc by the seat of the pants. This is certainly not an inherently implausible approach to artificial intelligence. The major hazard of work in this approach is that theories can become very nebulous, and that research can devolve into little more than the collection of striking anecdotes and the constructions of demonstration programs that work on a handful of examples.

5.1.3 Large-scale approaches

There have been a number of attempts to construct very large knowledge bases of commonsense knowledge by hand. The largest of these is the CYC program. This was initiated in 1984 by Doug Lenat, who has led the project throughout its existence. Its initial proposed methodology was to encode the knowledge in 400 sample articles in a one-volume desk encyclopedia together with all the implicit background knowledge that a reader would need to understand the articles (hence the name) (Lenat, Prakash, & Shepherd, 1985). It was initially planned as a ten-year project, but continues to this day. In the last decade, Cycorp has released steadily increasing portions of the knowledge base for public or research use. The most recent public version, OpenCyc 4.0, released in June 2012 contains 239,000 concepts and 2,039,000 facts, mostly taxonomic. ResearchCyc, which is available to be licenced for research purposes, contains 500,000 concepts and 5,000,000 facts.

A number of successful applications of CYC to AI tasks have been reported (Conesa, Storey, & Sugumaran, 2010) including web query expansion and refinement (Conesa, Storey, & Sugumaran, 2008), and question answering (Curtis, Matthews, & Baxter, 2005), and intelligence analysis (Forbus, Birnbaum, Wagner, Baker, & Witbrock, 2005).

CYC is often mentioned as a success of the knowledge-based approach to AI; for instance Dennett (2013) writes, "CYC is certainly the most impressive AI implementation of something like a language of thought" (p. 156). However, it is in fact very difficult for an outsider to determine what has been accomplished

here. In its first fifteen years, CYC published astonishingly little. Since about 2002, somewhat more has been published, but still very little, considering the size of the project. No systematic evaluation of the contents, capacities, and limitations of CYC has been published.[8]

It is not, for example, at all clear what fraction of CYC actually deals with commonsense inference, and what fraction deals with specialized applications such as medical records or terrorism. It is even less clear what fraction of commonsense knowledge of any kind is in CYC. The objective of representing 400 encyclopedia articles seems to have been silently abandoned at a fairly early date; this may have been a wise decision; but it would be interesting to know how close we are, 30 years and 239,000 concepts later, to achieving it; or, if this is not an reasonable measure, what has been accomplished in terms of commonsense reasoning by any other measure. There are not even very many specific *examples* of commonsense reasoning carried out by CYC that have been published.

There have been conflicting reports about the usability of CYC from outside scientists who have tried to work with it. Conesa, Story, and Sugumaran (2010) report that CYC is poorly organized and very difficult to use:

> The MT's [microtheory] Taxonomy in ResearchCyc is not very usable for several reasons:
> 1. There are over 20 thousand MT's in Cyc with the taxonomical structure of MT's being as deep as 50 levels in some domains.
> 2. There are many redundant subtype relationships that make it difficult to determine its taxonomical structure.
> 3. Some of the MT's are almost empty but difficult to discard.
> 4. Not all the MT's follow a standard representation of knowledge

They further report a large collection of usability problems including problems in understandability, learnability, portability, reliability, compliance with standards, and interface to other systems. More broadly, CYC has had comparatively little impact on AI research — much less than less sophisticated online resources as Wikipedia or WordNet.

5.2 Web mining

During the last decade, many projects have attempted to use web mining techniques to extract commonsense knowledge from web documents. A few notable examples, of many:

The KnowItAll program (Etzioni, et al., 2004) collected instances of categories by mining lists in texts. For instance, if the system reads a document containing a phrase like "pianists such as Evgeny Kissin, Yuja Wang, and Anastasia Gromoglasova" then the system can infer that these people are members of the category Pianist. If the system later enounters a text with the phrase "Yuja Wang, Anastasia Gromoglasova, and Lyubov Gromoglasova," it can infer that Lyubov Gromoglasova may also be a pianist. (This technique was first proposed by Marti Hearst (1992); hence patterns like "W's such as X,Y,Z" are known as "Hearst patterns".) More recently, the Probase system (Wu, Li, Wang, & Zhu, 2012), using similar techniques, has automatically compiled a taxonomy of 2.8 million concepts and 12 million isA relations, with 92% accuracy.

The NELL[9] (Never-Ending Language Learner) program (Mitchell, et al., 2015)has been steadily accumulating a knowledge base of facts since January 2010. These include relations between individuals, taxonomic relations between categories, and general rules about categories. As of January 2015, it has accumulated 89 million candidate facts, of which it has high confidence in about 2 million. However, the facts are of very uneven quality (table 3). The taxonomy created by NELL is much more accurate, but it is extremely lopsided. As of 6/9/2015 there are 9047 instances of "amphibian" but zero instances of "poem".

[8] A number of organizations have done private evaluations but the results were not published.
[9] http://rtw.ml.cmu.edu/rtw/

Fact	Confidence
federica_fontana is a director	91.5
illustrations_of_swollen_lymph_nodes is a lymph node	90.3
lake_triangle is a lake	100.0
louis_pasteur_and_robert_koch is a scientist	99.6
Illinois_governor_george_ryan is a politician	99.8
stephen is a person who moved to the state california	100.0
louis_armstrong is a musician who plays the trumpet	99.6
cbs_early_show is a company in the economic sector of news	93.8
knxv is a TV station in the city phoenix	100.0
broncos is a sports team that plays against new_york_jets	100.0

Table 3: Facts recently learned by NELL (6/11/2015)

Moreover, the knowledge collected in web mining systems tends to suffer from severe confusions and inconsistencies. For example in the Open Information Extraction system[10] (Etizoni, Fader, Christensen, Soderland, & Mausam, 2011) the query "What is made of wood?", (as of 6/9/2015) receives 258 answers of which the top twenty are: “The frame,” “the buildings”, “Furniture”, “The handle”, “Most houses”, “The body”, “the table”, “Chair”, “This one”, “The case”, “The structure”, “The board”, “the pieces”, “roof”, “toy”, “all”, “the set”, and “Arrow”. Though some of these are reasonable ("Furniture", "Chair"), some are hopelessly vague ("the pieces") and some are meaningless ("this one", "all). The query "What is located in wood?" gets the answers “The cemetery”, “The Best Western Willis”, “the cabins”, “The park” “The Lewis and Clark Law Review,” “this semi-wilderness camp”, “R&R Bayview”, “'s Voice School “[sic], and so on. Obviously, these answers are mostly useless. A more subtle error is that OIE does not distinguish between "wood" the material (the meaning of the answers to the first query) and "wood" meaning forest (the meaning of the answers to the second query).[11]

All of these programs are impressive — it is remarkable that you can get so far just relying on patterns of words, with almost no knowledge of larger-scale syntax, and no knowledge at all of semantics or of the relation of these words to external reality. Still, they seem unlikely to suffice for the kinds of commonsense reasoning discussed above.

5.3 Crowd sourcing

Some attempts have been made to use crowd-sourcing techniques to compile knowledge bases of commonsense knowledge (Havasi, Pustjekovsky, Speer, & Lieberman, 2009). Many interesting facts can be collected this way, and the worst of the problems that we have noted in web mining systems are avoided. . For example, the query "What is made of wood?" gets mostly reasonable answers; the top 10 are : “paper”, “stick”, “table”, “chair”, “pencil”, “tree”, “furniture”, “house”, “picture frame”, and “tree be plant and”,[12]

What one does not get, however, is the analysis of fundamental domains and the careful distinguishings of different meanings and categories needed to support reliable reasoning. For example, naïve users will not work out systematic theories of time and space; it will be difficult to get them to follow, systematically

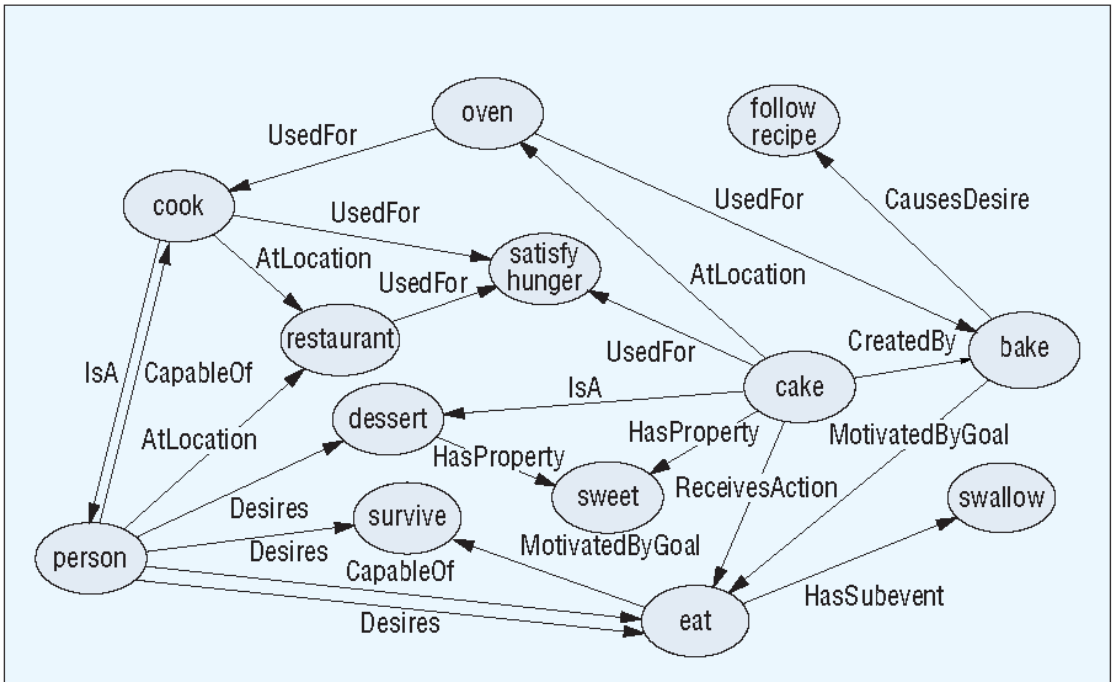
[10] http://openie.cs.washington.edu/

[11] Thanks to Leora Morgenstern for helpful discussions.

[12] This test was carried out in November 2014.

and reliably, theories that the system designers have worked out. As a result, facts get entered into the knowledge base without the critical distinctions needed for reasoning. Instead, one winds up with a bit of mess.

Consider, for example, the fragment of the crowd-sourced Concept Net 3.5 shown in figure 4 (Havasi, Pustjekovsky, Speer, & Lieberman, 2009).



. Figure 4: Concepts and relations in ConceptNet (from (Havasi, Pustjekovsky, Speer, & Lieberman, 2009))

Even in this small network, we see many of the kinds of inconsistencies most famously pointed out by Woods (1975). Some of these links always hold (e.g. “eat HasSubevent swallow”), some hold frequently (e.g. “person Desires eat”) and some only occasionally (e.g. “person AtLocation restaurant”). Some – like “cake AtLocation oven” and “cake ReceivesAction eat” – cannot be true simultaneously. The node “cook” is used to mean a profession in the link “cook isA person” and an activity in “oven UsedFor cook” (and in “person CapableOf cook”). Both cook and and cake are “UsedFor satisfy-hunger”, but in entirely different ways. (Imagine a robot who, in a well-intentioned effort at satisfying the hunger of its own owner, fricassees a cook.) On a technical side, the restriction to two-place relations also limits the expressivity in important ways. For example, there is a link “restaurant UsedFor satisfy-hunger”, another link might easily specify that, “restaurant UsedFor make-money”. But in this representational system there would be no way to specify the fact that the hunger satisfaction and money making have distinct beneficiaries (viz the customers versus the owner). All of this could be fixed post-hoc by professional knowledge engineers, but at enormous cost, and it is not clear whether crowds could be efficiently trained to do adequate work that would avoid these troubles.

6. Going Forward

We doubt that any silver bullet will easily solve all the problems of commonsense reasoning. As table 2 suggests, each of the existing approaches has distinctive merits, hence research in all these directions should presumably be continued. In addition, we would urge the following:

Benchmarks: There may be no single perfect set of benchmark problems, but as yet there is essentially none at all, nor anything like an agreed-upon evaluation metric; benchmarks and evaluation marks would serve to move the field forward.

Evaluating CYC. The field might well benefit if CYC were systematically described and evaluated. If CYC has solved some significant fraction of commonsense reasoning, then it is critical to know that, both as a useful tool, and as a starting point for further research. If CYC has run into difficulties, it would be useful to learn from the mistakes that were made. If CYC is entirely useless, then researchers can at least stop worrying about whether they are reinventing the wheel.

Integration. It is important to attempt to combine the strengths of the various approaches to AI. It would useful for instance to integrate a careful analysis of fundamental domains developed in a mathematically grounded theory with the interpretation of facts accumulated by a web mining program; or to see how facts gathered from web mining can constraint the development of mathematically grounded theories.

Alternative modes of reasoning. Neat theories of reasoning have tended to focus on essentially deductive reasoning (including deduction using default rules). Large-scale knowledge bases and web mining have focused on taxonomic and statistical reasoning. However, commonsense reasoning involves many different forms of reasoning including reasoning by analogy; frame-based reasoning in the sense of (Minsky, 1975); abstraction; conjecture; and reasoning to the best explanation. There is a substantial literature in some of these areas in cognitive science and informal AI approaches, but much remains to be done to integrate them with more mainstream approaches.

Cognitive science. Intelligent machines need not replicate human techniques, but a better understanding of human common sense reasoning might be a good place to start.

Acknowledgements

Thanks to Leora Morgenstern, Ken Forbus and William Jarrold for helpful feedback, to Ron Brachman for useful information, and to Thomas Wies for checking our German.

References

Allison, B., Guthrie, D., & Guthrie, L. (2006). Another Look at the Data Sparsity Problem. *Text, Speech, and Dialogue: 9th International Conference*, (pp. 327-334).

Baader, F., Horrocks, I., & Sattler, U. (2008). Descriptions Logics. In F. Van Harmelen, V. Lifschitz, & B. Porter, *Handbook of Knowledge Representation* (pp. 135-179). Amsterdam: Elsevier.

Bar-Hillel, Y. (1960). The present status of automatic translation of languges. In F. Alt, *Advances in Computers* (Vol. 1, pp. 91-163). New York: Academic Press.

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., & Sivic, J. (2014). Weakly supervised action labeling in videos under ordering constraints. *ECCV*, (pp. 628-643).

Brewka, G., Niemelli, I., & Truszczynski, M. (2008). Nonmonotonic Reasoning. In F. Van Harmelen, V. Lifschitz, & B. Porter, *Handbook of Knowledge Representation* (pp. 239-284). Amsterdam: Elsevier.

Cohn, A., & Renz, J. (2007). Qualitative Spatial Reasoning. In F. van Harmelen, V. Lifschitz, & B. Porter, *Handbook of Knowledge Representation* (pp. 551-596). Elsevier.

Conesa, J., Storey, V., & Sugumaran, V. (2008). Improving web-query processing through semantic knowledge. *Data and Knowledge Engineering, 66*(1), 18-34.

Conesa, J., Storey, V., & Sugumaran, V. (2010). Usability of upper level ontologies: The case of ResearchCyc. *Data and Knowledge Engineering, 69*, 343-356.

Curtis, J., Matthews, G., & Baxter, D. (2005). On the effective use of Cyc in a question answering system. *IJCAI Workshop on Knowledge and Reasoning for Answering Questions.*

Davis, E. (1998). The naive physics perplex. *AI Magazine, 19*(4), 51-79. From http://www.aaai.org/ojs/index.php/aimagazine/article/view/1424/1323

de Kleer, J. (1975). *Qualitative and quantitative knowledge in classical mechanics.* MIT AI Lab.

de Kleer, J., & Brown, J. (1985). A qualitative physics based on confluences. In D. Bobrow, *Qualitative Reasoning about Physical Systems* (pp. 7-84). Cambridge: MIT Press.

Dennett, D. (2013). *Intuition Pumps and Other Tools for Thinking.* Norton.

Etizoni, O., Fader, A., Christensen, J., Soderland, S., & Mausam. (2011). Open Information Extraction: The Second Generation. *IJAI*, (pp. 3-10).

Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popsecu, A., Shaked, T., . . . Yates, A. (2004). Web-scale extraction in KnowItAll (preliminary results). *13th International Conference on World Wide Web*, (pp. 100-110). Retrieved from http://dl.acm.org/citation.cfm?id=988687

Ferrein, A., Fritz, C., & Lakemeyer, G. (2005). Using Golog for Deliberation and Team Coordination in Robotic Soccer. *KI (Kuntzliche Intelligenz), 19*(1), 24-30.

Fisher, M. (2008). Temporal Representation and Reasoning. In F. Van Harmelen, V. Lifschitz, & B. Porter, *Handbook of Knowledge Representation* (pp. 513-550). Amsterdam: Elsevier.

Forbus, K. (1985). Qualitative process theory. In D. Bobrow, *Qualitative Reasoning about Physical Systems* (pp. 85-168). Cambridge, Mass.: MIT Press.

Forbus, K., Birnbaum, L., Wagner, E., Baker, J., & Witbrock, M. (2005). Analogy, Intelligent IR, and Knowledge Integration for Intelligence Analysis: Situation Tracking and the Whodunit Problem. *International Conference on Intelligence Analysis.*

Fromherz, M., Bobrow, D., & de Kleer, J. (2003). Model-based Computing for Design and Control of Reconfigurable Systems. *AI Magazine, 24*(4), 120.

Gene Ontology Consortium. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic acids research, 32*(suppl. 1), D258-D261.

Gentner, D., & Forbus, K. (2011). Computational models of analogy. *WIREs Cognitive Science, 2*, 266-276.

Halpern, J. (2003). *Reasoning about Uncertainty.* Cambridge, MA: MIT Press.

Havasi, C., Pustjekovsky, J., Speer, R., & Lieberman, H. (2009). Digital Intuition: Applying Common Sense Using Dimensionality Reduction. *IEEE Intelligent Systems, 24*(4), 24-35. doi:dx.doi.org/10.1109/MIS.2009.72

Hearst, M. (1992). Automatic acquisition of hyponyms from large text corpora. *ACL*, (pp. 539-545).

Hofstadter, D., & Sander, E. (2013). *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking.* New York: Basic Books.

Kalyanpur, A. (2012). Structured data and inference in DeepQA. *IBM Journal of Research and Development, 53*(3-4), 10:1-14.

Kolodner, J. (1993). *Case-based reasoning.* San Mateo, Calif.: Morgan Kaufmann.

Kuehne, S. (2004). *Understanding natural language description of physical phenomena.* Evanston, I.: Northwestern University.

Kuipers, B. (1986). Qualitative simulation. *Artificial Intelligence, 29*, 289-338.

Lenat, D., Prakash, M., & Shepherd, M. (1985). CYC: Using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks. *AI Magazine, 6*(4), 65-85.

Levesque, H., Davis, E., & Morgenstern, L. (2012). The Winograd schema challenge. *Principles of Knowledge Representation and Reasoing.*

Lovett, A., Tomei, E., Forbus, K., & Usher, J. (2009). Solving Geometric Analogy Problems through Two-Stage Analogical Mapping. *Cognitive Science, 33*(7), 1192-1231.

Miller, G. (1995). WordNet:A Lexical Database for English. *Communications of the ACM, 38*(11), 39-41.

Minsky, M. (1975). A Framework for Representing Knowledge. In P. Winston, *The Psychology of Computer Vision.* New York: McGraw Hill.

Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Betteridge, J., & Carlson, A. (2015). Never Ending Learning. *AAAI.*

Mueller, E. (2006). *Commonsense Reasoning.* San Francisco: Morgan Kaufmann.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* San Mateo, CA: Morgan Kaufmann.

Peppas, P. (2008). Belief Revision. In F. Van Harmelen, V. Lifschitz, & B. Porter, *Handbook of Knowledge Representation* (pp. 317-359). Amsterdam: Elsevier.

Pisanelli, D. (2004). *Ontologies in Medicine.* Amsterdam: IOS Press.

Price, C., Pugh, D., Wilson, M., & Snooke, N. (1995). The FLAME System: Automating electrical failure mode and effects analysis (FEMA). *IEEE Reliability and Maintainability Symposium*, (pp. 90-95).

Reiter, R. (2001). *Knowledge in Action: Logical Foundations for Specifying and Implementing Dynamical Systemsi.* Cambridge, Mass.: MIT Press.

Russell, B., Torralba, A., Murphy, K., & Freeman, W. (2008). Labelme: a database and web-based tool for image annotation. *International Journal of Computer Vision, 77*(1-3), 157-173.

Schank, R., & Abelson, R. (1977). *Scripts, Plans, Goals, and Understanding.* Hillsdale, N.J.: Lawrence Erlbaum.

Schubert, L. (2015). Semantic Representation. *AAAI.*

Shepard, B., Matuszek, C., Fraser, C., Wechtenhiser, W., Crabbe, D., Gungordu, Z. J., . . . Larson, E. (2005). A knowledge-base approach to network security: Applying Cyc in the domain of network risk assessment. *Association for the Advancement of Artificial Intelligence*, (pp. 1563-1568).

Surdeanu, M. (2013). Overview of the tac2013 knowledge base population evaluation: English slot filling and temporal slot filling. *Proceedings of the Sixth Text Analysis Conference (TAC 2013).*

Winograd, T. (1972). *Understanding Natural Language.* New York: Academic Press.

Woods, W. (1975). What's in a Link: Foundations for Semantic Networks. In D. Bobrow, & A. Collins, *Representation and Understanding: Studies in Cognitive Science.* New York: Academic Press.

Wu, W., Li, H., Wang, H., & Zhu, K. Q. (2012). Probase: A probabilistic taxonomy for text understanding. *ACM SIGMOD.*