The Pursuit of Machine Common Sense

By Joseph Churilla
Advisor: Chris Callison-Burch

EAS499 Senior Capstone Thesis

School of Engineering and Applied Science

University of Pennsylvania

April 29, 2020

Abstract

Despite widespread innovation in artificial intelligence (AI) over the past several decades, machines have yet to develop the ability to understand the basics of communication, physics, and psychology that humans take for granted, also known as "common sense". For machines to achieve human-like intelligence and move past the current niche-specific applications of AI, machines must develop this common sense/general intelligence. Previous approaches to achieving machine common sense (MCS) can be separated into the following categories: knowledge-based, web mining, and crowd sourcing [1]. Knowledge-based systems include mathematical methods (i.e. situation calculus [2], naïve physics [3], etc.), less formal methods (i.e. scripts [4], etc.) and large-scale approaches (i.e. logic ontologies such as Cyc [5], etc.) [1]. However, knowledge-based methods are limited by the fragility of codified, symbolic knowledge that fails to encompass the scope and subtlety of human common sense. On the other hand, web mining and crowd sourcing approaches (i.e. KnowItAll [6], NELL [7], etc.) are more efficiently scalable but fail to possess deep semantic understanding [8]. Going forward, researchers have proposed two categories of research strategies for achieving MCS in hope of further progress: first, constructing computational models that learn from experience, perhaps replicating a child's cognition for objects, agents and places ("bottom-up"); and second, constructing a common sense knowledge archive learned from reading the Web for answering natural language and image-based questions ("top-down") [8].

Introduction: What is Common Sense?

The goal of this literature review is to perform an in-depth analysis of previous research regarding machine common sense (MCS), to discuss currently developing research strategies for achieving MCS, and to explore future applications and benefits of developing artificial intelligence with MCS capabilities, Artificial General Intelligence (AGI). Artificial Intelligence is often used "as an umbrella term to describe the overall objective of making computers apply judgment as a human being would" [9]. Over the past several decades, technological innovation has intertwined artificial intelligence capabilities with day-to-day human life. Today's AI research can be roughly broken down into several distinct research areas: "Search and Optimization," "Fuzzy Systems", "Natural Language Processing and Knowledge Representation", "Computer Vision", "Machine Learning and Probabilistic Reasoning", and "Planning and Decision Making" [9]. Despite the numerous capabilities of AI across these areas, machine reasoning remains "narrow and highly-specialized", mandating cautious training and programming of every possible scenario [8].

Given these limitations, AI technologies, such as machine learning models, have been developed based upon historically observed data and thus often fail to account for unobserved data and its associated impact. Current AI functionality lacks what can be described as "human common sense" or "human general intelligence". The dictionary defines "common sense" as "sound and prudent judgment based on a simple perception of the situation or facts" [10].

Humans are gifted with the ability to retain and apply a certain level of knowledge/information about the world to unique, potentially unseen problems. This common, unspoken internal knowledge base is composed of "a general understanding of how the physical world works (i.e., intuitive physics); a basic understanding of human motives and behaviors (i.e., intuitive psychology); and knowledge of the common facts that an average adult possesses" [8]. For example, "common sense" tells us that, when one person states that he/she is traveling from New York to Paris tomorrow for vacation, the individual is likely traveling via plane rather than foot, bike, or boat. Without access to this uniquely human understanding of the surrounding world, machines are limited in their ability to perform basic human tasks such as rational decision-making under unique environmental conditions, learning from new situations, and communicating naturally with people [8]. Another elusive aspect of common sense is that it is multimodal (necessitating the use of several human senses), and thus does not fit well into existing AI research disciplines. "Common sense" is not solely learning from fewer examples, constructing a common knowledge database, or recognizing images/signals, but all these things working together in a joint system. As Davis and Marcus state, the "great irony of common sense" is that its information everyone knows but is unable to define it exactly or create machines with it [15].

Theory of Knowledge and its Representation

We can begin examining the literature on machine common sense by first understanding some of the classical theories surrounding knowledge and its representation by machines. The earliest work proposing commonsense representation by machines came in 1959 with John McCarthy's paper "Programs with Common Sense". McCarthy highlights the lack of progress in machine basic verbal reasoning processes that nearly any human can do quite easily [11]. This unique human ability, common sense, is achieved by a system that "automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows" [11]. Any programs aiming to achieve human-like intelligence ought to internally represent behaviors, express changes in these behaviors, have an improving mechanism for behavior that itself is improvable, be able to evolve and handle partial success on difficult problems, and lastly "create subroutines which can be included in procedures as units" [11]. Specifically, McCarthy proposes an "advice taker" program that solves problems via formal logic [11]. Rather than incorporating heuristics in the program, the "advice taker['s]" rules "will be described as much as possible in the language itself" [11]. McCarthy's representational language of choice is formal logic, likely predicate calculus, allowing a system to perform deduction from a set of logical premises. [11].

In 1973, Eugene Charniak presented a theory of knowledge for understanding basic texts such as children's stories, noting that an individual must have commonsense background knowledge to answer questions about them [12]. Charniak uses an example of a short story about a girl shaking a piggy bank (PB) and money falling out of it. Next, he poses questions such as

“Why did Janet get the PB?” and “Why was the PB shaken?” [12]. Although it is obvious to any human that one shakes a piggy bank to retrieve money, likely to use the money for some purpose, it is technically impossible to “deduce an answer from the statements in the story without using general knowledge about the world”, such as “shaking helps get money out of a PB” [12]. This example highlights the essential role commonsense knowledge plays in humans’ understanding of stories and the need for machines to have access to this knowledge if it is to achieve human-level narrative understanding. Furthermore, Charniak proposes that any such model would translate text into some internal representation composed of assertions, which would then “try to ‘fill in the blanks’ of a story on a line by line basis” [12]. To assist in the deduction process during question answering, the system would label the extracted facts and beliefs with “topic concepts,” and “fact finder” theorems could “establish facts which are comparatively unimportant” [12].

Despite these early attempts at formulating theories of knowledge, much less-optimistic research exists, suggesting that attempts to duplicate human cognition could be futile. McCarthy himself acknowledged the difficulties in understanding human nature and noted scientists’ trouble to explain how the great diversity of animal life can be expressed by small genetic variations, noting that “the problem of how such a representation controls the development of a fertilized egg into a mature animal is even more difficult” [11]. Others, such as Christopher Cherniak, state that “from a resource-realistic approach to cognitive science” a program of the human mind is impossibly cumbersome and unknowable: “The mind’s program would be an impossibility engine in that it would be practically unfeasible for us fully to comprehend and evaluate it” [13]. Any “computational approximation of the mind” would be massive, “branchy”, “quick-and-dirty”, and unpolished, implying that any such program is “fundamentally dissimilar to more familiar software” [13]. Similarly, Hubert Dreyfus argues that attempts at machine simulation of cognition “in digital computer language systematically excludes three fundamental human forms of information processing (fringe consciousness, essence/accident discrimination, and ambiguity tolerance)” [14]. For example, human and machine pattern recognition differ drastically. Humans can ignore noisy data and label it as unimportant and have an ability to “distinguish the essential from the inessential” [14]. Humans have “tolerance for changes in orientation”, imperfect and distorted information, and “background noise” that machines lack [14]. Moreover, humans “need not conceptualize or thematize” common traits to identify patterns, while machine recognition takes place “on the explicit conceptual level of class membership” [14]. Dreyfus also casts doubt on the “associationist assumption” that human thought can be broken down into easily understood, explicit processes, adding that even if the mind’s processes could be quantified into equation, there would be time constraints [14]. With the human brain still not fully comprehended, it clear that many differences between machines and humans and unknown unknowns pose further challenges for those attempting to duplicate human intelligence.

Although many researchers cast doubt on the manageability of replicating the human mind in full, the possibility of a much simpler structured program that could produce human intelligence is not ruled out [13]. In Cherniak's view, progress in AI applications was "outpacing theory" of how such programs worked [13]. In order to combat this incongruence, Cherniak suggests a shift in research focus to include "neuroanatomy and neurophysiology" and that "cognition, after all, is accomplished with a brain", paralleling today's research agenda of studying human cognition in infants (discussed later). Other theories suggest that a machine will only be able to achieve natural language understanding if it can "learn about the world," given the omitted facts and context-dependence comprehension requires [14]. Systems aiming to reach human cognition must be able to "distinguish the essential from the inessential features" of pattern instances, understand context clues, and "use cues which remain on the fringes of consciousness" [14].

In more recent commentary, Gary Marcus and Ernest Davis lay out a high-level theory for attaining deep understanding as follows: (i) begin with representing human core knowledge (discussed later), (ii) plant these representations into an architecture that can incorporate all types of knowledge, (iii) create reasoning processes that can handle complexity, incompleteness, and uncertainty, (iv) integrate these with "perception, manipulation, and language" to "build rich cognitive models of the world", and lastly (v) combine the AI's knowledge acquisition and reasoning features to build a human-like system that interacts with the outside world, learning and improving upon its existing knowledge just as in human development [15]. In order to begin, researchers must decide what type of knowledge is to be possessed by the intelligent machine and how this will be represented internally. Although much research in psychology has focused on developing frameworks for human's understanding of the world, few of the current big-data AI techniques have yet to take them into account [15]. Overall, although replicating the human mind from scratch seems implausible, creating systems with the ability to learn new information at a "conceptual and causal level" and learn theories beyond simple facts ought to lead to more promising, versatile and powerful AI [15].

Machine Common Sense Use Cases: Why is it important?

Today's artificial intelligence systems are overly sensitive and fragile. The development of machine common sense would allow researchers to apply AI technology beyond the current boundaries of niche environments. Such achievement would grant machines a deeper understanding and awareness of the world, human-like adaptability to the unexpected, and the ability to more naturally communicate with humans [8]. Some broad use cases for MCS that apply to AI systems are "sensemaking" (interpreting sensor/data from its external environment), checking the reasonableness of machine decision-making (monitoring actions and their safety in unfamiliar or new situations), "human-machine collaboration" (effective communication between machines and human users), and "transfer learning" (reapplying common sense for adaptation under new circumstances without hyper-specialized training) [8].

More specifically, other key areas that would benefit from MCS include natural language processing, computer vision, and robotics [1]. Consider the problem of machine translation and the lexical ambiguities humans can easily understand but persistently evade the grasp of software such as Google Translate. Translation software is often successful in using nearby words to predict simple translations with statistics, avoiding the task of true understanding [1]. However, when certain phrases and clauses are added between such words, the "statistical proxy for common sense" that was correct on the simple examples runs into problems as the complexity increases [1]. For example, as of 2015, Google Translate could handle the ambiguity of the word "working" when translating the following sentences into German: "The electrician is working" and "The telephone is working" [1]. However, the system used the German word for "laboring" when translating more complex sentences such as "The telephone on the desk is working" [1]. Although the statistical prowess of such systems will increase over time, perfect disambiguation will only realistically be achieved when systems attain true understanding of the text with the correct domain knowledge [1]. With respect to computer vision, similar challenges exist. Movies, for example, require the audience to piece together several different scenes, make inference about people's intentions, understand physical objects, relationships between characters, etc. [1]. Lastly, in robotics, if machines are to be trusted in uncontrolled environments, systems must be able to reasonably and safely handle unexpected situations without having been trained upon them. For example, an assistant robot asked to pour out a glass of wine by its owner ought to be able to think for itself and realize that if the wine glass pulled off the shelf is cracked, scratched, or dirty it ought to select another [1]. In other words, current robots are "literalists," that only perform based upon exact specifications and lack the flexibility of the human mind [15]. Only until machines can achieve the adaptability and reasonableness of human thought will robots be suited for complicated, open-ended environments (i.e. public stores, crowded streets, private homes, etc.), significantly increasing their usefulness in society [15].

What are the challenges to overcome?

Tackling Human-like Reasoning

The list of machine commonsense challenges remaining unsolved is numerous, spanning topics such as planning, physical and spatial reasoning, natural language understanding, intuitive psychology, etc. [16]. Certain reasoning challenges such as temporal reasoning, causality, and cross-domain understanding must be solved to achieve general intelligence capabilities and are relevant across most AI applications (vision, NLP, etc.). For example, robots instructed to perform some task (such as cooking a dinner) must understand the sequence of sub-tasks involved (i.e. mixing ingredients before placing in the oven). This sort of "temporal logic can allow the robot to construct a cognitive model" of events and combine the model with commonsense knowledge to develop a well-organized plan over time [15]. Understanding causality across time and space is also crucial, allowing machines to predict consequences of their actions and help formulate plans under novel circumstances [15]. Further, machines will need to be able to apply and combine learned knowledge across domains to solve real-world problems. Take healthcare robots, for example. If a hospital robot is to help elderly patients get into their beds, they must understand the patients' psychology, biology, the physics involved, and more to ensure safe completion of the task (i.e. just moving a patient's center of mass over a bed could result in great injury if the robot doesn't understand the orientation of the patient's body and how the patient is feeling) [15]. Although machines will need to have access to commonsense knowledge in order to tackle many of these challenges, this alone is not enough. Machines will need to possess human-like reasoning systems integrated with such commonsense knowledge in order to be of use in the dynamic, unsupervised real world.

Inherent Difficulties in Human Language

Although researchers have found success in image recognition techniques and other pattern matching systems, natural language processing has been more of a challenge due to the nature of human communication. Catherine Havasi explains that text is "precise and abbreviated", leaving out the boring basics in order to demonstrate creativity and make writing interesting [17]. For example, when one tells another about a trip to Starbucks, there's no need to explain that a hot drink comes in a mug or that one must pay a cashier before receiving the beverage. As such, much of the 'important glue" is left out of the data sets machines will encounter [17]. Simply put, humans are incentivized not to be boring: Grice's maxims suggest that humans generally only provide as much information as needed to stay on topic and relevant in a clear and concise manner but enough to avoid vagueness and ambiguity [17] [18]. Another troubling aspect of human language is what Geoffrey Nunberg calls "the social differentiation of knowledge": essentially a given word can mean different things to different people due to their unique backgrounds [19]. Nunberg poses the following questions: how does one determine what information a speaker associates with a word and, moreover, what is one able to infer about the speaker's "internal state" from this? [19]. In fixed domains without "social differentiation of knowledge" formal semantics can perform well, but without this constraint commonsense inference is crucial to language understanding.

Representation vs. Reasoning

Among the most important decisions researchers need to make is how best to represent knowledge internally so that programs can access it efficiently and reason with it effectively. Many of the classical theories suggest starting with formal, first-order logic [11] [20], but this approach is also widely critiqued [21] [22]. Robert Moore argued that many "important features of commonsense reasoning can be implemented only within a logical framework", such as those "involving incomplete knowledge of a problem situation" that require "deductive inference" [20]. Other problems requiring checking if "an existentially quantified proposition is true" and case-based reasoning all require some sort of formal logic [20]. Formal logic has no bound for what can be handled as an object, allowing for representation of the non-physical: “times, events, kinds, organizations, worlds, and sentences” [20]. Unrestricted by domains, first-order logic may always prove useful in dealing with reasoning about topics for which there is incomplete information [20].

Formal logic may be useful in deduction, but humans use a wide array of reasoning techniques that mathematical logic cannot handle as simply. Humans also rely on personal experience, analogy, anecdote, and probabilistic thinking, all of which logicists seem to ignore [21]. Some logicists even go to the extreme, claiming that inference does not generate new knowledge since it is already solved, which implies that logicism fails to characterize cross-domain knowledge application [21]. When applied to language, formal logic faces more challenges since meaning is context-dependent and often ambiguous. With the goal of logical forms to represent sentences more clearly than the text itself in a context-independent fashion, logical systems probably must maintain “distinct representations for the different reading of ambiguous natural-language expressions” [22]. Opposing logicism, Birnbaum calls for a “functional semantics” that only attributes meaning to words based on its use in practice, stating that, although knowledge and its use go hand-in-hand, knowledge from one application can be applied to several others [21].

More broadly, rather than fixating on formal logic as the representational language, there is a broader tradeoff between any machine’s internal representation of knowledge and its reasoning capabilities. As the expressivity of the chosen representational language increases, so do difficulties in tractability of reasoning [23]. To combat this issue, Levesque suggests to “push the computational barrier as far back as possible” and to loosen the “notion of correctness” [23]. Regardless of the chosen representational language, the knowledge representation system must have a compatible reasoning mechanism for question answering and the integration of new knowledge, reaching enough reliability in both resource-use and correctness [23].

Some researchers go so far as to hypothesize that “representation is the wrong unit of abstraction” [24]. When AI systems rely on interaction with the real world through sensory data and interaction, “reliance on representation disappears” [24]. Rather than diving deep into niche subproblems of AI, Rodney Brooks suggests building up intelligent creatures in layers and maintaining complete systems at each step to guarantee proper connectivity between sub-pieces [24]. Brooks critiques the decomposition of AI research into subfields such as “knowledge representation” and “qualitative reasoning”, claiming that human intelligence is not well understood enough to decompose into the correct sub-systems, let alone to tie them all together [24]. Overall, Brooks places doubt on researchers’ ability to determine the exact requirements of a truly intelligent system and concludes that research should “use the world as its own model” [24]. Along a similar vein, Dan Roth proposes a “Learning to Reason” framework that assumes no exact knowledge representation for the system but rather that the system would create its own as it interacts with its external environment [25]. Roth presents the example of a “baby robot, starting out its life" and compares it to a human infant, stating that "nature would have provided for the infant a safe environment in which it can spend an initial period of time" to interact and learn from its environment before being expected to have "full functionality" [25]. This approach avoids placing constraints on the knowledge representation language and emphasizes the importance of integrating reasoning with learning from the start [25].

Limits of Deep Learning

Although deep learning has allowed for progress in areas like image recognition and helped machines beat experts at board games, the question arises whether these systems are just powerful statistical engines and glorified pattern matchers or if they actually understand and reason about topics as humans do. Unfortunately, the state-of-the-art neural systems fit more into the former category. Even the best deep neural networks can be fooled by images, confidently labeling certain unrecognizable pixels completely wrong (i.e. mistaking "white noise static" for a lion) [26]. Appendix Figure 1 provides several examples of these "fooling images" that are unidentifiable to the human eye but predicted with over 99% certainty to be discernable objects [26]. Nevertheless, there is hope that future neural networks will be furnished with commonsense, causal reasoning, intuitive physics, and other features that will substantially improve their capabilities [27]. Brenden Lake et al. suggest incorporating "more structure and inductive biases" into neural models to attain more human-like learning [27]. Future neural nets could also be programed to "learn to learn" to generalize knowledge across domains, make better inferences with less training data, and avoid starting from scratch [27]. In summary, current machine-learning typically aims at taking some hyper-specific, niche task and tries to "bootstrap it from scratch" [15], but, without generalization and the ability to build up on previously learned knowledge, achieving general artificial intelligence is unlikely.

Classical Benchmarks for Achieving Common Sense

In 1950, Alan Turing proposed an "imitation game," now widely known as the "Turing Test," to replace the question "can machines think?" [28]. The game is simple and involves two humans and a machine. One human is the interrogator, who asks the other human and the machine a series of questions and after some time must decipher which respondent is the machine and which is the other human. If the interrogator is unable to consistently distinguish between the human and computer, it is believed that the computer can think like a human [29]. However, this classical test depends upon unstructured conversation that can "facilitate deception and trickery" [29].

More recently in 2011, Hector Levesque proposed the Winograd Schema Challenge as a substitute for the Turing Test that avoids these problems [29]. This test is composed of a series of binary choice reading comprehension questions, with each question involving a pair of individuals or items and a pronoun/ possessive adjective referencing one of the parties that is also suitable for the other object [29] [30]. Consider the following example: “The town councillors refused to give the angry demonstrators a permit because they feared violence. Who feared violence? Answer 0: the town councillors Answer 1: the angry demonstrators” [29]. Levesque explains that in order to solve this problem one must possess and reason with background knowledge, a process that necessitates thinking [29]. To humans, the set of schemas are so easily solvable that the answerer may not even realize the ambiguity; however, there is “no obvious statistical test over text” that will provide consistent accuracy in its answers [30]. A list of other currently outstanding, relevantly applied commonsense reasoning problems exists on Stanford’s “Common Sense Problem Page” [16].

Attempts at Machine Common Sense To-Date

Previous approaches to achieving machine common sense (MCS) can be separated into the following categories: knowledge-based, web mining, and crowd sourcing [1] (See Appendix Figure 2 for a visual representation). Knowledge-based systems include mathematical methods (i.e. situation calculus [2], naïve physics [3], etc.), less formal methods (i.e. scripts [4], etc.) and large-scale approaches (i.e. logic ontologies such as Cyc [5], etc.) [1]. Below, each of these categories is described in more detail.

Knowledge-based Systems

Knowledge-based attempts at commonsense reasoning involve creating representations to handle different types of knowledge and reasoning with these representations [1]. The earliest example of such attempts is attributed to AI pioneer John McCarthy who suggested using formal logic to handle commonsense reasoning [1] [11]. Since then, researchers have pursued a variety of formal and informal logic-based frameworks and logic-based ontologies [8].

i. Mathematical and Logic-Based Attempts

While several “technically demanding” mathematical frameworks for common sense reasoning have been developed, little work has been done to implement these “purely theoretical” and “technically demanding” foundations into practice [1]. One such example is situation calculus, which employs first-order logic to model states and actions. Often used in planning, situation calculus utilizes a “branching model of time” and analyzes alternative actions [1] but fails to be of use in areas such as narrative understanding since “it treats events as atomic” and mandates knowledge of event sequence [1]. An example of situation calculus is STRIPS (STanford Research Institute Problem Solver) that uses large sets of predicate calculus formulas as a world model and “employs a resolution theorem prover to answer questions of particular models and uses means-ends analysis to guide it to the desired goal-satisfying model” [2].

Logical frameworks have also been formulated to handle commonsense challenges such as default reasoning [31], circumscription and non-monotonic reasoning [32], and descriptions [33]. Commonsense reasoning in humans frequently includes default reasoning when making inferences with incomplete information that are later subject to change by later observation [1]. These inferences are often of the form "in the absence of any information to the contrary, assume…" [31]. With negative conjectures greatly outnumbering the positive about any given environment, some claim that only positive information need be specified as negative inferences are assumed by default (i.e. the closed-world assumption) [31]. Similarly, McCarthy proposed first-order-logic for dealing with circumscription, the conjecture rule humans often use when "jumping to certain conclusion" [32]. For example, a "qualification problem" arises when someone is to complete a certain task (such as taking a rowboat across a river): there are nearly infinite qualifications that could be formulated that must be satisfied for success task completion (the oars are in the boat and well-functioning) [32]. Circumscription makes concrete the informal human assumption that a "tool can be used for its intended purpose unless it can't" [32]. Additional work has gone into developing logical methods for expressing concepts and the relationships between them ("description logics") [1]. For example, KL-ONE is a knowledge representation system that supplies "a language for expressing an explicit set of beliefs for a rational agent" [33]. Lastly, logic has also been theorized to be able to express domains such as naïve physics [3]. Most famously, Pat Hayes proposed a theory for formalizing all of naïve physics "in a declarative symbolic form" that would be broken into concept clusters such as "forces and movements," "energy and effort," and many more [34].

ii. Informal Attempts

Other less formal knowledge-based approaches have contributed to the field by theorizing about a wide array of specific inference techniques [1]. One notable attempt is Minsky's concept of frames [36]. Minsky argues that mathematical logic focuses too heavily on consistency and thus ignores that human thought starts with "suggestive but defective plans and images" that are modified over time [36]. Frames are proposed as a type of "data structure for representing a stereotyped situation" that humans pull from their memory and adapt to novel circumstances to reason about attributes of and expectations for a given event [36]. This framework for inference seems quite relevant to commonsense thought as humans draw upon templates for certain events (i.e. a football game) and actions (i.e. entering a kitchen) to consider characteristics and potential outcomes of situations (routine and novel) in everyday life.

Schank's theory of scripts [4] parallels the idea of frames but is specifically applied to the "case of structured collections of events" [1]. Additionally, Schank proposed the concept of plans that uses commonsense knowledge to order events so that individuals can achieve some goal [4]. This theory has proven useful in not only understand the attributes of certain situations but also how complex behavior is organized, assisting inference about unique situations and the complexities human behavior more broadly [36].

Furthermore, there has been mild success in developing frameworks for qualitative reasoning about the physical world [37]. More generally, qualitative reasonings involves "direction of change in interrelated quantities" such as the relationships between price and demand of a product [1]. However, most work in this area has succeeded in specific domains but as complexity increases these systems lack usefulness and cannot generalize across domains [1].

Of note, several modern software applications, such as "text editors, operating systems shells, and so on," are all rather informal in nature and do not rely on deeply mathematical and statistical architecture [1]. Thus, informal attempts at commonsense AI are not necessarily far-fetched if application catches up to theory.

iii. Large-scale Attempts

The last broad category of knowledge-based approaches to comes in the form of massive databases containing what humans would classify as commonsense information (i.e. cars drive on roads or the sky is blue). In 1984, Doug Lenat started building Cyc [5], one of the largest attempts at a logic ontology to-date. Lenat proposes Cyc "as an expert system with a domain that spans all everyday objects and actions" [5]. Most of Cyc's commonsense assertions are obvious to any human: "you have to be awake to eat" or "you can usually see people's noses, but not their hearts" [5]. The Cyc system structures its assertions around a variety of contexts, creating an organizational architecture "reminiscent of Schankian scripts" [5]. Rather than focusing on a solution to general intelligence, Lenat explains that Cyc was developed to "build a set of micro-theories that together cover the common cases of each problem", sacrificing full inference functionality for "expressiveness and efficiency" [5].

The underlying idea behind Cyc is that machines need the information at hand not only to complete assigned tasks but also predict future actions, which requires vast amounts of world knowledge [38]. Several proposed applications of Cyc include functionality as a "semantic backbone" to connect data sources, improving text editors with advancements in grammar checking, and enhancing the authenticity of objects and agents in simulations [5]. Nevertheless, after several decades of work, critics claim that the proposed benefits of Cyc have failed to come to fruition [15]. Little information about the true contents of the system have been made public and usability concerns have arisen regarding the dependability of the system and its interfacing with other applications [1].

Over the last several decades, a collection of more specialized ontologies has arisen for specific domains. For example, researchers created WordNet as a lexical database for applications in linguistics and natural language processing [39]. WordNet is a thesaurus-like network in which "sets of synonym form the basic building blocks" [40]. In order to build off of WordNet's focus on semantic relations, researchers later created VerbNet as verb-centric lexical database that integrate with WordNet but that also make use of syntactic information [41].

As previously noted, one of Cyc's criticisms is its public unavailability. In 2001, researchers with the goal of creating an easier-to-use version of Cyc introduce the SUMO ontology, "merging publicly available ontological content into a single, comprehensive, and cohesive structure" [42]. Next, the YAGO ontology emerged as an enhanced version of WordNet that pulled common information about the world from Wikipedia [43]. Other notable contributions include DOLCE [44] to further refine WordNet and the proposal of the Semantic Web which aimed to use ontologies to improve the World Wide Web.

Over the past decades, there have been numerous attempts at constructing logic ontologies built upon commonsense facts, but still there is criticism that having access to such information is not enough for true comprehension of "how these 'facts connect'" [15]. Machines ought to not only record particular facts but also incorporate them into a broader framework. For example, rather than just knowing that Picasso painted *The Old Guitarist* or that Beethoven composed the "Ode to Joy," systems ought to fit observations "into a larger framework that makes clear that a creator owns a work until he/she sells it, that works by a single person are often stylistically similar and so on" [15]. Ontologies tend to define knowledge in "black or white symbols, which never quite match the subtleties of human concepts they are intended to represent" [8]. Moreover, there are even further challenges in usability for many of the ontologies that developers seek to interface with [8]. Researchers aiming to use Cyc for web query expansion report several issues spanning API failures, redundancy in content, and insufficient information in certain domains, etc. [46]. Overall, logic ontologies have successfully been created in specific realms, but much progress remains in order to reach widespread applicability and any deeper grasp of human-like commonsense.

Web mining

Another broad category of commonsense research comes in the form of machine learning techniques implemented to search the Web for and extract information [8]. Web mining could allow for a more scalable approach to commonsense database creation as it is automated, as opposed to the manual input required to create many previous ontologies. One such automated information extraction program is KnowItAll, which collects information from text based off various syntactic patterns it can recognize [6]. For example, KnowItAll can successfully gather "instances of categories by mining lists in texts" [1]. KnowItAll is composed of an "extractor" that handles the rules relating to the syntactic patterns, a "search engine interface" that creates queries, an "assessor" that utilizes a naïve Bayes classifier to produce an estimate of correctness for the extracted knowledge, and lastly stores all of this in a database [6]. DBpedia [47] is another extraction program aimed at generating structured data from Wikipedia and other datasets online in hopes to spur progress towards further development of the Semantic Web.

Additional progress has been made in the field with the creation of NELL (never-ending learning language) [8] [48]. The NELL program runs 24/7 on the Web and has accumulated a massive knowledge base of “confidence-weighted beliefs” [8]. Its developers also use crowdsourced feedback for quality checks and system improvement [8]. The system not only extracts semantic categorization and relationship information but also learns to improve its extraction over time [48]. NELL has even learned to extract knowledge from untraditional structures, like tables and lists, and to use “probabilistic horn clause rules” to make additional inferences from previously collected knowledge [48]. As opposed to common supervised learning models, NELL’s program is based off semi-supervised learning that “couple[s] the training of many different learning tasks” so that the program can learn “thousands of functions from only a small amount of supervision,” resembling a more human-like learning procedure [8]. Regardless, NELL is not flawless. The program has trouble self-monitoring progress, contains certain immutable methods unable to self-improve, and has no ability to handle temporal and spatial aspects of knowledge [8]. The outputted taxonomy is also reportedly quite lopsided with vast amounts of information in some domains but nearly none in other [1].

In summary, web-mining has shown promising improvements in efficiency over manual data accumulation, but these methods have yet to improve upon their “relatively shallow semantic representations,” limiting their usefulness in reasoning beyond basic question answering via querying through data [8]. Like logic ontologies, web-mining approaches have limited understanding of their contents and a shortfall of any human-like reasoning capabilities.

Crowd Sourcing via the Web

The last subcategory for attempts at commonsense reasoning is crowdsourcing commonsense facts from individuals. The idea is that since essentially all humans have this knowledge, commonsense knowledge can be extracted from them. Launched in 1999, the Open Mind Common Sense project is the first of its kind to implement crowdsourcing on the web to collect commonsense knowledge [49]. Open Mind uses a variety of crowdsourcing methods, including giving users a short story and then asking the user for text input about implied information [50]. An example might be stating “Bob had a cold. Bob went to the doctor” with the user responding “Bob was feeling sick” or “The doctor made Bob feel better” [50]. Rather than separating knowledge into siloed microtheries as in Cyc, Open Mind users are only required “to build topic vectors”, which are groupings of “concepts that are related to a given topic” [50]. Constructed over the output of the Open Mind project is the ConceptNet knowledge base [51]. Rather than focusing on “lexical categorization” like WordNet or “formalized logical reasoning” like Cyc, ConceptNet aims to provide “practical context-based inferences” [51]. The semantic network is filled with more in-depth relations, such “EffectOf,” “DesireOf,” and “CapableOf,” and can better handle complexities in natural language (analogy, space, time, ambiguity, etc.) that other attempts cannot [51].

A drawback of crowdsourcing is that collected data can be partial or inaccurate [15]. Even though all people understand commonsense knowledge, relying on individuals to input data in a form useful for machines can lead to inconsistencies and confusion [15]. Crowdsourced knowledge also lacks the "analysis of fundamental domains" and the differentiation between various meanings that are required for well-grounded reasoning [1].

Recent Approaches to Solving MCS

In Fall 2018, DARPA announced a new program to provide funding to researchers with the goal of tackling the "elusive" Machine Common Sense problem [8]. The program outlines four new major research areas ripe for breakthroughs: (i) new representations, (ii) commonsense extraction and the web, (iii) experiential learning, and (iv) replicating childhood cognition [8]. The corresponding research is highlighted in the following sections.

New Representations

As described earlier, classical attempts with mathematical logic and other knowledge-based attempts at commonsense reasoning have failed to truly encompass the wide array of reasoning capabilities possessed by humans. Recent research has focused on developing new representations more suitable for such complexities. One promising trajectory of research in natural language processing and computer vision is the rise of semantic embeddings. Neural nets can now use word embeddings, mappings of natural language words into numeric vectors, to better analyze semantic similarities of words in text based upon their proximity to other words throughout large corpora [8]. Most famously, Google's Word2Vec [52] [54] software is the state-of-the-art for word embedding creation [53]. The software uses the extremely efficient "Skip-gram model" whose objective is to create word embeddings that better predict nearby words in text [52]. The underlying hypothesis of the program is that "words in similar contexts have similar meanings", however this has yet to be proven theoretically rigorous [53]. Researchers also found that adding word vectors together and tokenizing phrases allows for better representations of longer texts [52]. Further development of the model has been used to create paragraph vectors, which have proven for more meaningful than bag-of-words representations that completely disregard word order and semantics [54]. With these powerful representations, Google has developed a neural "zero-shot translation" model for machine translation between multiple languages that can successfully translate language pairs absent from the training data [55]. This is significant because it exhibits that neural models can attain human-like transfer learning.

Another success in new language representations is knowledge enhanced embeddings (KEE) [56]. These embeddings "combine context and commonsense knowledge" and have helped researchers make significant improvements on the Winograd Schema challenge. Trained over commonsense "cause-effect word pairs" and large amounts of text, the KEE framework treats commonsense information as a "semantic constraint" [56]. Further refinement in integrating both semantic and commonsense constraints into embeddings appears promising for future improvement in natural language understanding.

Other new representations come in the form of probabilistic models of learning. Despite lack of true understanding of the human mind’s inner workings, Tenenbaum et al. suggest a “Bayesian approach” to machine inference to move the field closer [57]. Probabilistic learning avoids the “either-or dichotomies” that have divided AI researcher for decades [57]. As apposed to previous knowledge-based approaches, a system based on probability avoids problems related to the rigidity of structured logic and can handle “noisy data of experience” [57]. In 2015, Lake et al. proposed the “Bayesian program learning (BPL) framework” that exhibits human-like generalization abilities and learn “a large class of visual concepts” from a minimal number of examples [58].

In computer vision, researchers have made progress using And-Or graphs as a basis for a “stochastic and context sensitive” image grammar [59]. This works aims to solve the problems of handling several hundred “object and scene categories” while simultaneously allowing for “intra-category structural variation” [59]. The proposed grammar benefits from the recursive structure of the graphs and is integrated with a probabilistic framework [59]. This grammar has been applied to videos to extract commonsense knowledge and predict semantics of a given scene [60]. Even further, the grammar was applied within a framework to analyze text and video jointly, and was able to produce “narrative text descriptions” and answer basic questions (i.e. “who, what, when, where and why”) about the multimedia [61].

Lastly, researchers have continued to create new and improved knowledge bases for commonsense reasoning. One such example is ATOMIC, which is a knowledge graph that focuses on inferential (i.e. if-then) relationships between concepts to better capture human-like inference capabilities [62]. With training on vast amounts of crowdsourced inferential data in the form of free-form text, neural networks are given a certain scenario and can infer about previous events that happened leading up to it [62]. Although the ability to infer cause and effect relationships is crucial to commonsense reasoning, this is only one aspect of human reasoning machines will need to acquire if they are to achieve general intelligence.

Commonsense Extraction and the Web

New research has continued progress in the fields of web-mining and information extraction from text, picture, and video data [8]. Similar to NELL but for images rather than text, the NEIL system traverses the web and extracts commonsense knowledge by analyzing images [63]. Additionally, NEIL can label object categories, scenes, and their attributes [63]. In text extraction, researchers have created a “hunting framework” for commonsense information that was demonstrated to achieve progress on the Winograd Schema Challenge [64]. This framework processes each Winograd Schema question, automatically creates relevant search queries on the web, extracts the text, and uses the information to predict an answer [64]. Both of the above approaches represent successes in reasoning about commonsense via extraction from text and image data. The following three sub-sections detail more specific historical and recent approaches to such extraction and understanding techniques with text, vision, and robotics data.

i. Commonsense and Natural Language

Classical attempts at commonsense reasoning with natural language begin with the concept of semantic networks [67]. In 1975, W.A. Woods expressed the need to develop such a representation for natural language that can handle many practical problems inherent in natural language such as “relative clauses,” “degrees of uncertainty,” “times and tense” and many others [67]. In 1976, John McCarthy proposed the idea of an “artificial natural language” based in predicate calculus that could represent assertions made in text and assist in natural language understanding [67], but formal logic has widely proven too rigid to handle these complexities in an efficient manner. Over time, researchers highlighted many more challenges intrinsic in text. In 1985, Mooney and Dejong proposed a new narrative understanding framework for understanding different agents’ actions from their goals, which was more broadly an attempt at extracting causal inferences from text [68]. Other research, noting that text is often “vague, insufficient, and ambiguous,” introduced the concept of naïve semantics, which is the general commonsense knowledge every natural language speaker possesses [69]. From the above examples, one can tell that early research focused on many unique challenges to understanding natural language and gains appreciation for wide array of difficulties current researchers face when trying to extract information from and reason over bodies of text.

Recent research in commonsense knowledge extraction from and understanding of text is also scattered across various sub-challenges. Some have focused on identifying semantically plausible events from text by training models on crowdsourced data [70]. For example, commonsense reasoning tells us that a human could literally “swallow a paintball,” but this is likely not something training data would provide information on [70]. Others have focused on new ways to dynamically incorporate common-sense knowledge from ConceptNet and Wikipedia into neural natural language understanding models so that additional background knowledge can be drawn upon beyond the “static” information collected during training [71]. Moreover, other extraction techniques have centered around basic “object-property comparisons” by comparing word embeddings (i.e. “elephant and tiger”) of the compared objects to the embeddings of the compared qualities (i.e. “big and small”) [72]. Commonsense physical knowledge has also been extracted via inference from verbs via Verb Physics techniques [73]. For example, one knows that if someone enters a building that the building is bigger than the individual [73]. Further, researchers have extracted commonsense inferences “about the mental states of people in relation to events” [74], created “temporally aware” embeddings to incorporate a time dimension into entity relations [75], and trained models to extract commonsense information for action justification [77]. Finally, researchers have also created models of state changes to better comprehend the ordering of procedural text [78] and frameworks for understanding and reasoning about people's emotional reactions to events [79] (See Appendix Figure 3 for an annotated example of this emotion/motivation framework).

ii. *Commonsense and Vision*

As image data is now more widely available than ever, computer scientists have turned to visual data to assist in the commonsense extraction process. Part of commonsense knowledge involves understanding one's physical surroundings and being able to reason about them. Researchers have turned to image and video data to start chipping away at this task. A very basic starting point is reasoning about physical characteristics of objects, such as their size. Some research has begun by building models that take both textual and image data as inputs to learn the relative sizes of different objects in the scenes [80]. Other strategies focus on extracting spatial relationships between object (i.e. "'holds(bed,dog)'" and "'laying-on(bed,dog)'" [81]) to detect new commonsense relations (i.e. "'holds(furniture, domestic animal)"') [81]. Additional research on spatial knowledge extraction has also been developed to learn spatial relationships from annotated images where the spatial information is implicit (i.e. "'man riding horse'") [82]. In addition, reasoning has been implemented in programs that use vision and text data to reason about the plausibility of commonsense assertions [83]. Furthermore, commonsense not only involves understanding of general physical relationships but also reasoning about the effects of certain actions on physical objects. AI systems will need to "understand basic action-effect relations" about real world items if they are to ever collaborate with humans [84]. To assist in this effort, models have learned to (i) given an input image, predict actions that likely occurred beforehand and (ii) given an action in the form of a verb-noun pair, predict the resulting image (See Appendix Figure 4 for some examples) [84]. In summary, several pieces of recent commonsense research in computer vision have shown that using both text and image data greatly improved learning ability and enabled better commonsense extractions about the physical world that text-only systems are less likely to grasp.

iii. Commonsense and Robotics

Every day, humans receive a variety of sensory inputs (sound, sight, touch, etc.) and make commonsense judgements with them. Programming robots to collect some of this same sensory information could further enhance machine intelligence efforts. Arguing that physical interaction is a crucial component of learning, some computer scientists have built robots with visual and touch sensors to learn from physical interaction directly [85]. Rather than just passively observing images and videos, the robot "pushes, pokes, grasps, and observes objects in a tabletop environment" and train a ConvNet with the sensor data for image classification [85]. Other technologies have been developed to turn videos into semantically labeled 3D scenes [86]. The building of 3D simulation environments has also allowed for the creation of new commonsense reasoning tests, such as "Embodied Question Answering" in which "an agent is spawned at a random location in a 3D environment and asked a question" about an item in its surroundings [87].

Commonsense Extraction and the Web (cont.)

Rather than creating new representations for extracting commonsense from scratch, several researchers have focused on knowledge base completion (KBC) to enhance the use of current commonsense knowledge bases. One example of a KBC success involved training a model to give quality ratings to novel ConceptNet tuples [88] [89]. Other improvements have come in the field of deep neural networks. Some of these examples include augmenting neural networks with "explicit memory" [90]. These "memory networks" have access to a "long-term memory" that "effectively acts as a (dynamic) knowledge base" [91]. Additionally, some new neural network functionality has been developed to "understand procedural text through (neural) simulation of action dynamics" [92]. Applied in the "cooking domain," for example, the model successfully used entity embeddings and learned "action transformations" that resulted in modifications of the state of entities [92].

Experiential Learning

A key aspect of human learning is learning from experience. Across several different domains, humans make inferences about the future based on passed observations. Recent work in computer vision has made headway in duplicating this sort of learning. One significant advancement came in a 2016 that used video footage to train models to infer what could occur in the future [93]. Unlabeled video data proves very useful in training these models because they are readily available at low cost and highly scalable [93]. Rather than collecting pixel data from each frame, the study extracted semantic "visual representations" [93]. The study exploits the "temporal structure in unlabeled video" to "anticipate human actions and objects" [93]. This is an important breakthrough because event prediction technology has many applications in practice, including integration with product recommendation and security systems [93]. Another hypothesis is that from this technology, machines could eventually better understand human behavior and psychology, making them more suitable collaborators with humans.

Other applications for experiential learning are found in the realm of basic physics. Facebook AI researchers recently completed a study on physical relationships in the real world in which they trained models to observe falling stacks of wooden blocks [95]. Humans do not need to understand complicated mathematical formulae to comprehend elementary physics concepts such as gravity, rather humans "rely on intuition, built up through interaction with the real world" [95]. As such, in this study, the researchers trained convolutional neural networks to predict if blocks would fall and, if they would, what the trajectory of each block would be [95]. The research team notes that the learned models could also generalize to novel situations with additional added blocks and pictures of real blocks (instead of the 3D models used in training) and still achieve human-level prediction accuracy [95]. Although this is just a beginning, future work could revolve around training models on other physical attributes and combining work to create a system that truly understands human-level intuitive physics.

Replicating Childhood Cognition

Classical theorists throughout history have brought up the idea of trying to simulate human cognition in order to achieve artificial general intelligence. In 1950, Turing stated, “Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain” [28]. However, until over the past few decades, there was little understanding of cognitive development of humans from infancy to adulthood, let alone the inner working and the human brain (still largely unknown). Humans remain superior to machines in several reasoning tasks, which has led to a movement to reverse engineer the human’s cognitive development.

This view is seemingly supported by the Theory of Grounded Cognition [96] [97]. Grounded cognition theorists believe that “modal simulations, bodily states, and situated action underlie cognition” [96] and assume “no central module for cognition” [97] (See Appendix Figure 5 for a visualized explanation of grounded cognition [97]). Gallese and Lackoff hypothesize that “understanding of concrete concepts—physical actions, physical objects, and so on—requires sensory motor simulation” [98], hinting at value in duplicating human cognition processes in machines. Additionally, Lackoff argues that metaphor is essential to human though as humans “typically conceptualize the nonphysical in terms of the physical” [99]. When considering where to begin, computer scientists have turned to developmental psychologists’ Theory of Core Knowledge [100] [101] [102]. The Theory of Core knowledge states that children are “endowed with several distinct core systems of knowledge,” including “objects, agents, number, and space” [100]. These distinct core systems later serve as foundations for more powerful cognitive abilities [101]. For example, one study of preschoolers found that “language understanding is intertwined with commonsense psychology” [105]. By reviewing the literature in developmental psychology and better understanding the timeline by which human infants reach certain milestones in their cognitive processes, researchers can better understand the milestones and benchmarks machines must accomplish to replicate the development of machine cognition.

Some of the most recent research in this field builds upon the notion that children develop expectations for objects, agents, and places and act “surprise when those principles are violated” (i.e. VOE) [8]. One psychological study in this area claims that children’s learning is stimulated by VOE events [103]. VOE events encourage young children analyze the defying object’s properties further and “test relevant hypotheses for that object’s behavior” [103]. Building off this revelation, researchers at MIT Early Childhood Lab have started crowdsourcing babies' reactions to VOE events and recording them on video to analyze the infants' facial reactions [8] [104].

A further attempt at replicating human cognition involved probabilistic representation of human thought [106]. Battaglia et al. proposed an "intuitive physics engine" model to make predictions about physical objects in scenes with incomplete information [106]. The study relies on the idea that individuals run quick mental simulations in their head to generate inferences about physical scenes and that thus a probabilistic program could be trained to predict in a similar fashion [106]. Another study attempted to integrate the VOE concept into deep learning models [107]. The study was trained on videos simulations and proved plausible the viability of using deep learning "to extract fundamental physical principles" [107]. In summary, several approaches have found large breakthroughs in commonsense reasoning via reverse engineering human cognition, encouraging further collaboration between computer science and development psychology.

Evaluation of New Commonsense Techniques

DARPA highlights two main strategies for achieving a commonsense service: (i) replicating the experiential learning of human children with focus on core domains of knowledge and (ii) building a web-mining system with adult-like commonsense knowledge and reasoning capabilities [8]. The first approach will be benchmarked against the key domains of objects, agents, and places (See Appendix Figure 6 [8]). For example, systems will have to understand the basics of object motion, how agents can affect an object's motion, and navigate surfaces [8]. The second approach will be tested against The Allen Institute for Artificial Intelligence's crowdsourced commonsense test questions found in the SWAG dataset [108].

Conclusion

Achieving artificial general intelligence has been a goal of the AI community since its founding in the 1950s by John McCarthy. From mathematical theories muddled in technical minutiae to ontologies filled with millions of facts, no fully developed system has yet been able to parallel the capabilities of the human mind. Recent advancement in deep learning have successfully learned to translate languages and make predictions about specific physical reasoning tasks: but are these systems really exhibiting human-like reasoning or are they just glorified, statistics-powered machines? Will commonsense ontologies and networks ever be able to truly understand their contents, or are they just complex lists of facts? It is too early to determine. What we do know is that common sense is multimodal (necessitating the use of several human senses), and thus does not fit well into existing AI research disciplines. As such, a well-integrated approach across domains, inspired by work in developmental psychology, and focused on interactions with the real world appears most promising.

Appendix

Figure 1: Flawed labeling of evolved images by state-of-the-art deep neural nets [26]

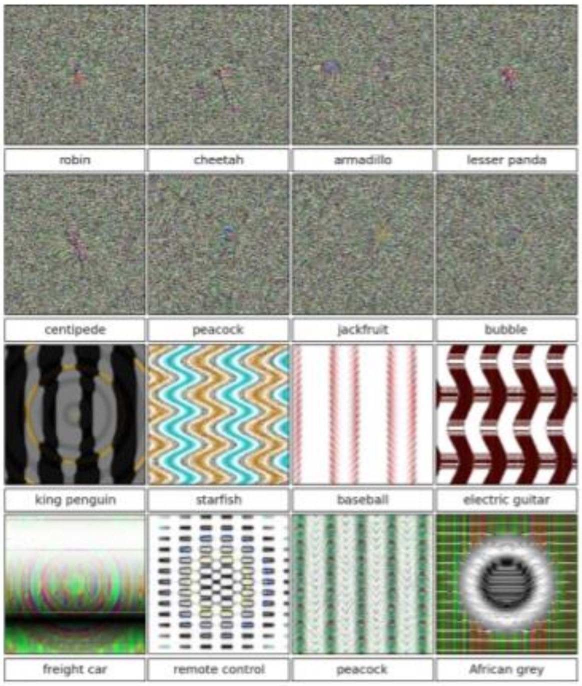


Figure 2: Overview of Historical Attempts at Commonsense Reasoning [1]

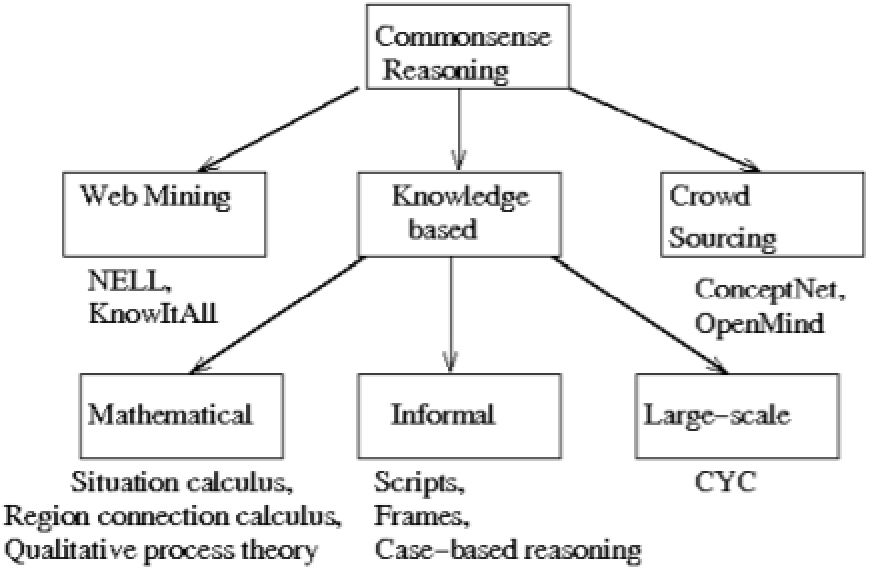
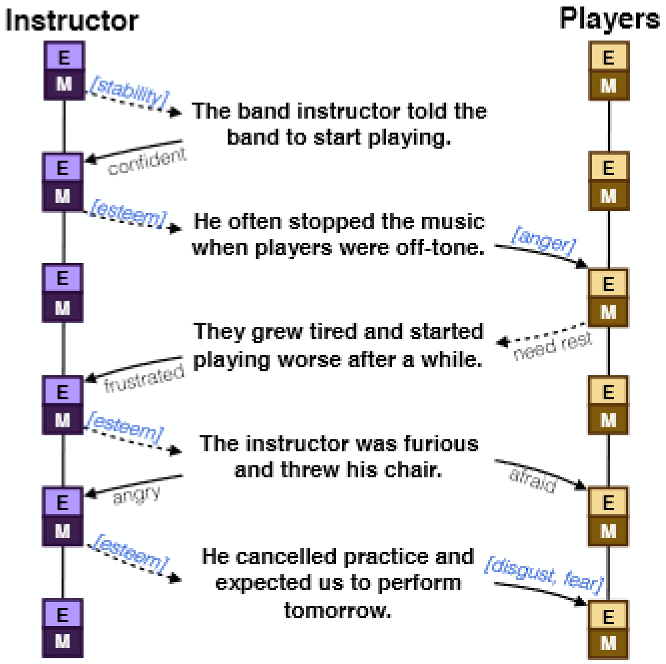


Figure 3: Annotated Example of Emotion and Motivation Extraction Framework from Text [79]



The band instructor told the band to start playing.

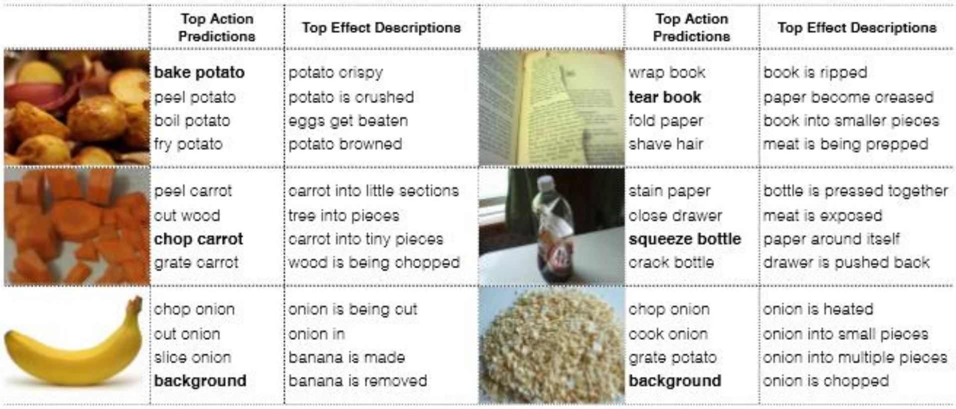
He often stopped the music when players were off-tone.

They grew tired and started playing worse after a while.

The instructor was furious and threw his chair.

He cancelled practice and expected us to perform tomorrow.

Figure 4: Action-Effect Inference Examples [84]



	Top Action Predictions	Top Effect Descriptions		Top Action Predictions	Top Effect Descriptions
	bake potato peel potato boil potato fry potato	potato crispy potato is crushed eggs get beaten potato browned		wrap book **tear book** fold paper shave hair	book is ripped paper become creased book into smaller pieces meat is being prepped
	peel carrot cut wood **chop carrot** grate carrot	carrot into little sections tree into pieces carrot into tiny pieces wood is being chopped		stain paper close drawer **squeeze bottle** crack bottle	bottle is pressed together meat is exposed paper around itself drawer is pushed back
	chop onion cut onion slice onion **background**	onion is being cut onion in banana is made banana is removed		chop onion cook onion grate potato **background**	onion is heated onion into small pieces onion into multiple pieces onion is chopped

Figure 5: Grounded Cognition Visualization [97]

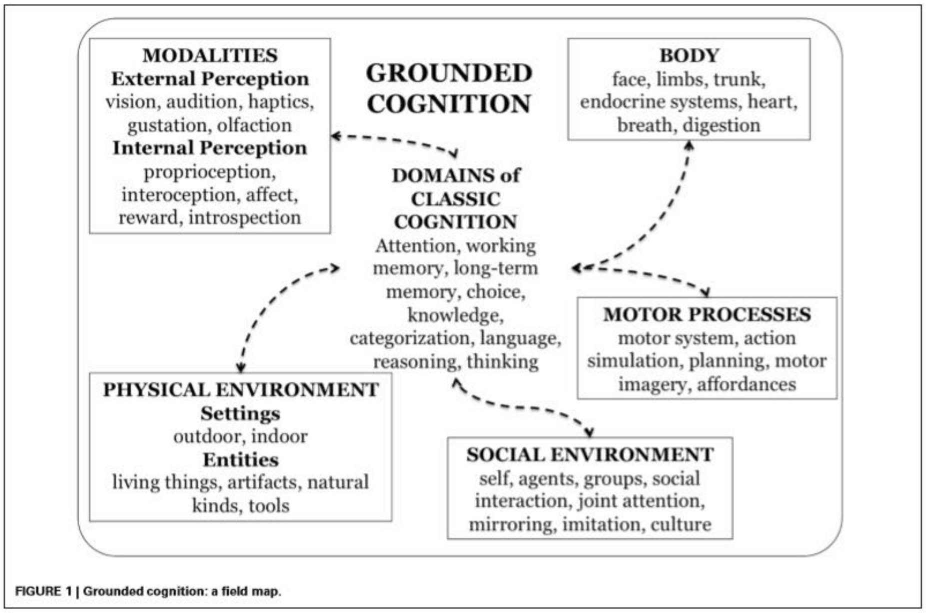
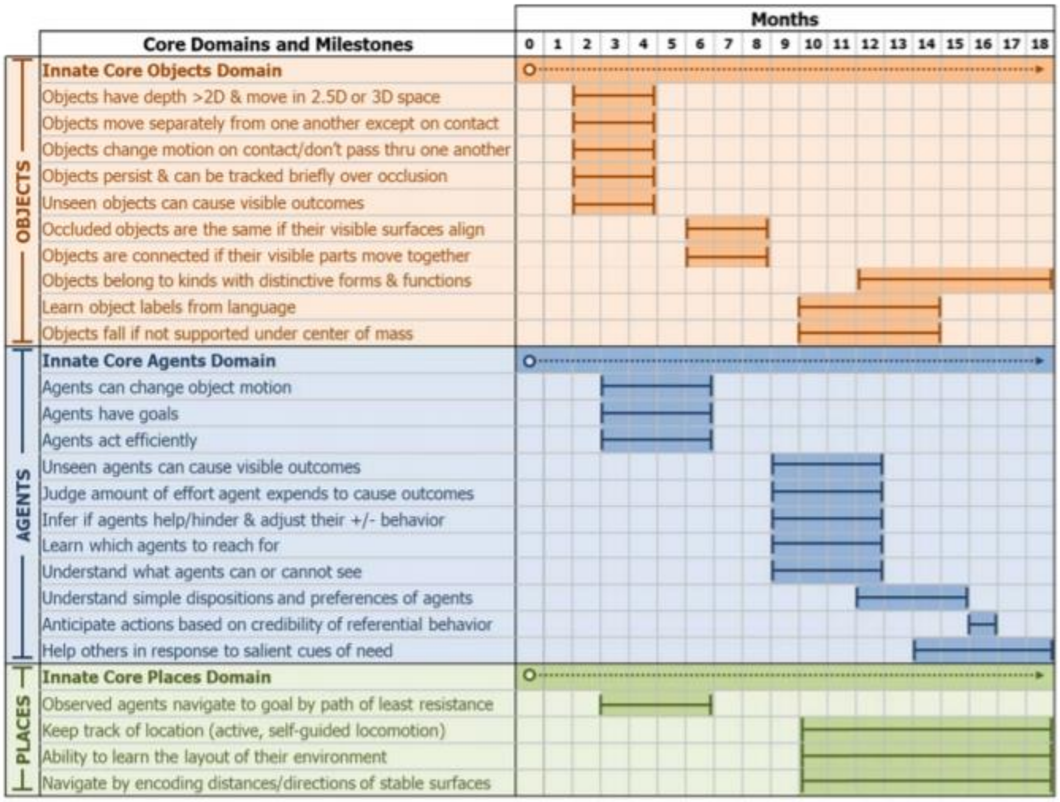

GROUNDED COGNITION
MODALITIES
External Perception
vision, audition, haptics, gustation, olfaction
Internal Perception
proprioception, interoception, affect, reward, introspection
BODY
face, limbs, trunk, endocrine systems, heart, breath, digestion
DOMAINS of CLASSIC COGNITION
Attention, working memory, long-term memory, choice, knowledge, categorization, language, reasoning, thinking
MOTOR PROCESSES
motor system, action simulation, planning, motor imagery, affordances
PHYSICAL ENVIRONMENT
Settings
outdoor, indoor
Entities
living things, artifacts, natural kinds, tools
SOCIAL ENVIRONMENT
self, agents, groups, social interaction, joint attention, mirroring, imitation, culture

FIGURE 1 | Grounded cognition: a field map.

Figure 6: Cognitive Development Milestones [8]



		Months																		
	Core Domains and Milestones	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
OBJECTS	Innate Core Objects Domain	○	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	▸
	Objects have depth >2D & move in 2.5D or 3D space			■	■	■														
	Objects move separately from one another except on contact			■	■	■														
	Objects change motion on contact/don't pass thru one another			■	■	■														
	Objects persist & can be tracked briefly over occlusion			■	■	■														
	Unseen objects can cause visible outcomes			■	■	■														
	Occluded objects are the same if their visible surfaces align							■	■	■										
	Objects are connected if their visible parts move together							■	■	■										
	Objects belong to kinds with distinctive forms & functions													■	■	■	■	■	■	■
	Learn object labels from language											■	■	■	■	■				
	Objects fall if not supported under center of mass											■	■	■	■	■				
AGENTS	Innate Core Agents Domain	○	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	▸
	Agents can change object motion				■	■	■	■												
	Agents have goals				■	■	■	■												
	Agents act efficiently				■	■	■	■												
	Unseen agents can cause visible outcomes										■	■	■	■						
	Judge amount of effort agent expends to cause outcomes										■	■	■	■						
	Infer if agents help/hinder & adjust their +/- behavior										■	■	■	■						
	Learn which agents to reach for										■	■	■	■						
	Understand what agents can or cannot see										■	■	■	■						
	Understand simple dispositions and preferences of agents													■	■	■	■			
	Anticipate actions based on credibility of referential behavior																	■		
	Help others in response to salient cues of need															■	■	■	■	■
PLACES	Innate Core Places Domain	○	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	▸
	Observed agents navigate to goal by path of least resistance				■	■	■	■												
	Keep track of location (active, self-guided locomotion)											■	■	■	■	■	■	■	■	■
	Ability to learn the layout of their environment											■	■	■	■	■	■	■	■	■
	Navigate by encoding distances/directions of stable surfaces											■	■	■	■	■	■	■	■	■

References

[1 Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92-103.

[2] Fikes, R. E., & Nilsson, N. J. (1971). STRIPS: A new approach to the application of theorem proving to problem solving. Artificial intelligence, 2(3-4), 189-208.

[3] Hayes, P. J. (1978). The naive physics manifesto. Institut pour les études sémantiques et cognitives.

[4] Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals and understanding: An inquiry into human knowledge structures. Lawrence Erlbaum.

[5] Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. Communications of the ACM, 38(11), 33-38.

[6] Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., ... & Yates, A. (2004, May). Web-scale information extraction in knowitall: (preliminary results). In *Proceedings of the 13th international conference on World Wide Web* (pp. 100-110).

[7] Mitchell, T., Cohen, W., Hruschka, E., Talukdar, P., Yang, B., Betteridge, J., ... & Krishnamurthy, J. (2018). Never-ending learning. *Communications of the ACM*, *61*(5), 103-115.

[8] Gunning, D. (2018). Machine common sense concept paper. *arXiv preprint arXiv:1810.07528*.

[9] Annoni, A., Cesar, R. M., Anzai, Y., Hall, W., Hardman, L., van Harmelen, F., ... & Keene, P. (2018). ArtificiaI Intelligence: How Knowledge Is Created, Transferred, and Used.

[10] “Common Sense.” *Merriam-Webster*, Merriam-Webster,w ww.merriamwebster.com/dictionary/common%20sense?src=search-dict-box.

[11] McCarthy, John. (1959). “Programs with Common Sense.”

[12] Charniak, E. (1973, August). Jack and Janet in Search of a Theory of Knowledge. In *IJCAI* (pp. 337-343).

[13] Cherniak, C. (1988). Undebuggability and cognitive science. *Communications of the ACM*, *31*(4), 402-412.

[14] Dreyfus, H. L. (1965). Alchemy and Artificial Intelligence. Santa Monica, CA: RAND Corporation.

[15] Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon.

[16] Common Sense Problem Page. Retrieved from www-formal.stanford.edu/leora/commonsense/.

[17] Havasi, C. Luminoso Technologies. (2017, April 18). *How to teach machines common sense*. YouTube. www.youtube.com/watch?v=G_ChiYI_9hs

[18] Schiffman, H. (2005) Grice's Maxims. Retrieved from www.sas.upenn.edu/~haroldfs/dravling/grice.html.

[19] Nunberg, G. (1987). Position paper on common-sense and formal semantics. In *Theoretical Issues in Natural Language Processing 3*.

[20] Moore, R. C. (1982). *The role of logic in knowledge representation and commonsense reasoning* (pp. 428-433). SRI International. Artificial Intelligence Center.

[21] Birnbaum, L. (1991). Rigor mortis: a response to Nilsson's “Logic and artificial intelligence”. *Artificial Intelligence, 47*(1-3), 57-77.

[22] Robert C. Moore. (1981). Problems in logical form. In *Proceedings of the 19th annual meeting on Association for Computational Linguistics* (ACL ’81). Association for Computational Linguistics, USA, 117–124.

[23] Levesque, H.J. & Brachman R.J. (1985). “A Fundamental Tradeoff in Knowledge Representation and Reasoning (Revised Version).

[24] Brooks, R. A. (1991). Intelligence without representation. *Artificial intelligence,* 47(1), 139-159

[25] Roth, D. (1996, November). Learning in order to reason: The approach. In *International Conference on Current Trends in Theory and Practice of Computer Science* (pp. 113-124). Springer, Berlin, Heidelberg.

[26] Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 427-436).

[27] Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences, 40*.

[28] Turing, A.M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433.

[29] Levesque, Hector J. (2011) “The Winograd Schema Challenge.” In *Proc. of the CommonSense-11 Symposium*.

[30] Davis, E., Morgenstern L., & Ortiz C. "The Winograd Schema Challenge." Retrieved from https://cs.nyu.edu/faculty/davise/papers/WinogradSchemas/WS.html

[31] Reiter, R. (1980). A logic for default reasoning. *Artificial intelligence, 13*(1-2), 81-132.

[32] McCarthy, J. (1981). Circumscription—a form of non-monotonic reasoning. In Readings in Artificial Intelligence (pp. 466-472).

[33] Brachman, R. J., & Schmolze, J. G. (1988). An overview of the KL-ONE knowledge representation system. In Readings in Artificial Intelligence and Databases (pp. 207-230).

[34] Davis, E. (1998). The naive physics perplex. AI Magazine, 19(4), 51-79.

[35] Minsky, M. (1974). A framework for representing knowledge.

[36] Schank, R. C., & Abelson, R. P. (1989). An early work in cognitive science. *Cognitive science*.

[37] Bobrow, D. G. (1984). Qualitative reasoning about physical systems: an introduction. *Artificial intelligence*, *24*(1-3), 1-5.

[38] Panton, K., Matuszek, C., Lenat, D., Schneider, D., Witbrock, M., Siegel, N., & Shepard, B. (2006). Common sense reasoning–from Cyc to intelligent assistant. In *Ambient Intelligence in Everyday Life* (pp. 1-31). Springer, Berlin, Heidelberg.

[39] WordNet. Retrieved from https://wordnet.princeton.edu/

[40] Miller, G. A. (1998). *WordNet: An electronic lexical database*. MIT press.

[41] Schuler, K. K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon.

[42] Niles, I., & Pease, A. (2001, October). Towards a standard upper ontology. In *Proceedings of the international conference on Formal Ontology in Information Systems-Volume 2001* (pp. 2-9).

[43] Suchanek, F. M., Kasneci, G., & Weikum, G. (2007, May). Yago: a core of semantic knowledge. In Proceedings of the 16th international conference on World Wide Web (pp. 697-706). ACM.

[44] Gangemi, A., Guarino, N., Masolo, C., Oltramari, A., & Schneider, L. (2002, October). Sweetening ontologies with DOLCE. In *International Conference on Knowledge Engineering and Knowledge Management* (pp. 166-181). Springer, Berlin, Heidelberg.

[45] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 284(5), 34-43.

[46] Conesa, J., Storey, V. C., & Sugumaran, V. (2010). Usability of upper level ontologies: The case of ResearchCyc. Data & Knowledge Engineering, 69(4), 343-356.

[47] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). Dbpedia: A nucleus for a web of open data. In *The semantic web* (pp. 722-735). Springer, Berlin, Heidelberg.

[48] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka, E. R., & Mitchell, T. M. (2010, July). Toward an architecture for never-ending language learning. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*.

[49] Common sense for artificial intelligence. Retrieved from https://www.media.mit.edu/projects/conceptnet-new/overview/

[50] Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., & Zhu, W. L. (2002, October). Open Mind Common Sense: Knowledge acquisition from the general public. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 1223-1237). Springer, Berlin, Heidelberg.

[51] Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, *22*(4), 211-226.

[52] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

[53] Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

[54] Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

[55] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Hughes, M. (2017). Google’s multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, *5*, 339-351.

[56] Liu, Q., Jiang, H., Ling, Z. H., Zhu, X., Wei, S., & Hu, Y. (2017, March). Combing context and commonsense knowledge through neural networks for solving winograd schema problems. In *2017 AAAI Spring Symposium Series*.

[57] Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, 331(6022), 1279-1285.

[58] Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332-1338.

[59] Zhu, S. C., & Mumford, D. (2006). A stochastic grammar of images. Foundations and Trends® in Computer Graphics and Vision, 2(4), 259-362.

[60] Si, Z., Pei, M., Yao, B., & Zhu, S. C. (2011, November). Unsupervised learning of event and-or grammar and semantics from video. In *2011 International Conference on Computer Vision* (pp. 41-48). IEEE.

[61] Tu, K., Meng, M., Lee, M. W., Choe, T. E., & Zhu, S. C. (2014). Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, *21*(2), 42-70.

[62] Sap, M., Le Bras, R., Allaway, E., Bhagavatula, C., Lourie, N., Rashkin, H., ... & Choi, Y. (2019, July). Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 33, pp. 3027-3035).

[63] Chen, X., Shrivastava, A., & Gupta, A. (2013). Neil: Extracting visual knowledge from web data. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1409-1416).

[64] Emami, A., De La Cruz, N., Trischler, A., Suleman, K., & Cheung, J. C. K. (2018). A knowledge hunting framework for common sense reasoning. *arXiv preprint arXiv:1810.01375*.

[66] Woods, W. A. (1975). What's in a link: Foundations for semantic networks. In *Representation and understanding* (pp. 35-82). Morgan Kaufmann.

[67] McCarthy, J. (1990). An example for natural language understanding and the AI problems it raises. *Formalizing Common Sense: Papers by John McCarthy*, *355*.

[68] Mooney, R. J., & DeJong, G. (1985, August). Learning schemata for natural language processing. In *IJCAI* (pp. 681-687).

[69] Dahlgren, K., McDowell, J., & Stabler, E. P. (1989). Knowledge representation for commonsense reasoning with text. *Computational linguistics*, *15*(3), 149-170.

[70] Wang, S., Durrett, G., & Erk, K. (2018). Modeling semantic plausibility by injecting world knowledge. *arXiv preprint arXiv:1804.00619*.

[71] Weissenborn, D., Kočiský, T., & Dyer, C. (2017). Dynamic integration of background knowledge in neural NLU systems. *arXiv preprint arXiv:1706.02596*.

[72] Yang, Y., Birnbaum, L., Wang, J. P., & Downey, D. (2018, July). Extracting commonsense properties from embeddings with limited human guidance. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 644-649).

[73] Forbes, M., & Choi, Y. (2017). Verb physics: Relative physical knowledge of actions and objects. *arXiv preprint arXiv:1706.03799*.

[74] Rashkin, H., Sap, M., Allaway, E., Smith, N. A., & Choi, Y. (2018). Event2mind: Commonsense inference on events, intents, and reactions. *arXiv preprint arXiv:1805.06939*.

[75] Dasgupta, S. S., Ray, S. N., & Talukdar, P. (2018). Hyte: Hyperplane-based temporally aware knowledge graph embedding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2001-2011).

[76] Collell, G., Van Gool, L., & Moens, M. F. (2018, April). Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[77] Yang, S., Gao, Q., Saba-Sadiya, S., & Chai, J. (2018). Commonsense Justification for Action Explanation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 2627-2637).

[78] Tandon, N., Mishra, B. D., Grus, J., Yih, W. T., Bosselut, A., & Clark, P. (2018). Reasoning about actions and state changes by injecting commonsense knowledge. *arXiv preprint arXiv:1808.10012*.

[79] Rashkin, H., Bosselut, A., Sap, M., Knight, K., & Choi, Y. (2018). Modeling naive psychology of characters in simple commonsense stories. *arXiv preprint arXiv:1805.06533*.

[80] Bagherinezhad, H., Hajishirzi, H., Choi, Y., & Farhadi, A. (2016, March). Are elephants bigger than butterflies? reasoning about sizes of objects. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[81] Yatskar, M., Ordonez, V., & Farhadi, A. (2016, June). Stating the obvious: Extracting visual common sense knowledge. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 193-198).

[82] Collell, G., Van Gool, L., & Moens, M. F. (2018, April). Acquiring common sense spatial knowledge through implicit spatial templates. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[83] Lin, X., & Parikh, D. (2015). Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2984-2993).

[84] Gao, Q., Yang, S., Chai, J., & Vanderwende, L. (2018, July). What action causes this? towards naive physical action-effect prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 934-945).

[85] Pinto, L., Gandhi, D., Han, Y., Park, Y. L., & Gupta, A. (2016, October). The curious robot: Learning visual representations via physical interactions. In *European Conference on Computer Vision* (pp. 3-18). Springer, Cham.

[86] Xiang, Y., & Fox, D. (2017). Da-rnn: Semantic mapping with data associated recurrent neural networks. *arXiv preprint arXiv:1703.03098*.

[87] Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., & Batra, D. (2018). Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 2054-2063).

[88] Li, X., Taheri, A., Tu, L., & Gimpel, K. (2016, August). Commonsense knowledge base completion. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1445-1455).

[89] Jastrzębski, S., Bahdanau, D., Hosseini, S., Noukhovitch, M., Bengio, Y., & Cheung, J. C. K. (2018). Commonsense mining as knowledge base completion? A study on the impact of novelty. *arXiv preprint arXiv:1804.09259*.

[90] Sukhbaatar, S., Weston, J., & Fergus, R. (2015). End-to-end memory networks. In *Advances in neural information processing systems* (pp. 2440-2448).

[91] Weston, J., Chopra, S., & Bordes, A. (2014). Memory networks. *arXiv preprint arXiv:1410.3916*.

[92] Bosselut, A., Levy, O., Holtzman, A., Ennis, C., Fox, D., & Choi, Y. (2017). Simulating action dynamics with neural process networks. *arXiv preprint arXiv:1711.05313*.

[93] Vondrick, C., Pirsiavash, H., & Torralba, A. (2016). Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 98-106).

[94] Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327-18332.

[95] Lerer, A., Gross, S., & Fergus, R. (2016). Learning physical intuition of block towers by example. *arXiv preprint arXiv:1603.01312*.

[96] Barsalou, L. W. (2008). Grounded cognition. Annual Review of Psychology, 59, 617-645.

[97] Pezzulo, G., Barsalou, L. W., Cangelosi, A., Fischer, M. H., McRae, K., & Spivey, M. (2013). Computational grounded cognition: a new alliance between grounded cognition and computational modeling. *Frontiers in psychology*, *3*, 612.

[98] Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive neuropsychology*, *22*(3-4), 455-479.

[99] Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

[100] Carey, S., & Spelke, E. (1996). Science and core knowledge. *Philosophy of science*, *63*(4), 515-533.

[101] Spelke, E. S. (2000). Core knowledge. *American psychologist, 55*(11), 1233.

[102] Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental science*, 10(1), 89-96.

[103] Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91-94.

[104] Schulz, L. 'Your Baby, the Physicist': Study Overview. Received from lookit.mit.edu/studies/cfddb63f-12e9-4e62-abd1-47534d6c4dd2/.

[105] Jara-Ettinger, J., Floyd, S., Huey, H., Tenenbaum, J. B., & Schulz, L. E. (2019). Social Pragmatics: Preschoolers Rely on Commonsense Psychology to Resolve Referential Underspecification. *Child development*.

[106] Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327-18332.

[107] Piloto, L., Weinstein, A., TB, D., Ahuja, A., Mirza, M., Wayne, G., ... & Botvinick, M. (2018). Probing physics knowledge using tools from developmental psychology. *arXiv preprint arXiv:1804.01128*.

[108] Zellers, R., Bisk, Y., Schwartz, R., & Choi, Y. (2018). Swag: A large-scale adversarial dataset for grounded commonsense inference. *arXiv preprint arXiv:1808.05326*.