Preprint. To appear in *Prometheus - Critical studies in Innovation*, issue TBD, 2023.

Review: Machines Like Us: Toward AI with Common Sense, by Ronald J. Brachman and Hector J. Levesque

Henry Kautz
henry.kautz@gmail.com

Since its birth in the 1940s, the field of artificial intelligence has been divided into two camps, one focused on artificial neural networks and the other on reasoning with symbolic representations of knowledge. The symbolic representational approach firmly dominated the field until 2012, when a neural network named AlexNet handily won an algorithm competition for recognizing objects in images [Krizhevsky et al. 2012]. Further convincing successes of neural network algorithms for speech recognition, the game of Go [Silver et al. 2017], and other problems that had long eluded the KR approach soon followed. Today, neural networks, under the banner of "deep learning", where "deep" refers to the fact that the artificial neurons are arranged in many layers, dominate research and commercial applications. Most students studying AI learn little about knowledge representation, and the approach is rarely mentioned in news stories and popular accounts of AI.

It would be a grave error, however, to conclude that the more than fifty years of research in knowledge representation and reasoning yielded no insights about the nature of computational intelligence. We are beginning to see deep learning researchers struggling with issues that have long been studied in knowledge representation, such as the nature of the concepts and categories that an AI system must employ to make sense of the world - an issue sometimes identified as "learning disentangled representations" [Bengio et al. 2013].

Furthermore, even today's most sophisticated deep learning systems are unreliable, in that they can unexpectedly and catastrophically fail on certain inputs. For example, at the time of writing this review, OpenAI's ChatGPT is the most powerful natural language processing system ever created, and its fluency in producing text has led some people to wrongly conclude that it must be conscious. It is not difficult, however, to find examples when ChatGPT produces plausible sounding but entirely erroneous "hallucinated" answers [Hofstadter 2022]. One path to trustworthy AI may be to base systems on explicit and validated representations of what they know and do not know [Marcus & Davis 2019]. It is therefore timely that Machine Like Us: Toward AI with Common Sense [Brachman & Levesque 2022] has appeared to remind us of the insights developed by the knowledge representation research community and to restate fundamental open problems that cannot even be stated without reference to how an intelligent being represents and reasons about the world.

Ronald Brachman and Hector Levesque are two preeminent researchers in the knowledge representation and reasoning tradition, individually and jointly publishing many influential papers from the 1980s through the present day. Brachman is best known for showing how so-called "semantic networks", a graph-based knowledge representation approach that had been

developed in the earliest days of AI and which is still in widespread use including by Google and Facebook under the name of "knowledge graphs", could be viewed as object-oriented versions of first-order logic [Brachman and Schmolze 1985]. Levesque collaborated on some of this work and went on to develop logic-based methods for reasoning about beliefs and plans [Levesque et al. 1997]. In 2012, Levesque and two other collaborators, Ernie Davis and Leora Morgenstern, introduced Winograd Schemas, a way to create challenge problems for natural language processing systems that could only be disambiguated by using commonsense knowledge; these problems have become a part of standard benchmark sets for such systems [Levesque et al. 2012].

The book begins by defining what they mean by commonsense: "... the ability to make effective use of ordinary, everyday, experiential knowledge in achieving ordinary, everyday, practical goals." They note that is related to but more specialized than rationality, which puts no restrictions on the complexity of the task or the amount of reasoning required. They further distinguish commonsense knowledge, "facts, patterns, principles, and generalizations", from commonsense reasoning, "the ability to make use of that knowledge in certain ways." These two themes of representation and reasoning are entwined throughout the book, although they can be separated; we shall have more to say on this separation below. There follows a summary of the history of knowledge representation in AI, from John McCarthy's 1958 paper on "Programs with Common Sense" [McCarthy 1958], which proposed using first-order logic to represent knowledge and automated deduction for reasoning, through the era of expert systems in the 1970s and 1980s, to today's emphasis on deep learning.

We then come to what I consider to be the most valuable part of the book, the three chapters that discuss what it means to represent knowledge (Chapter 4), a general classification or ontology of kinds of knowledge (chapter 5), and a further categorization of kinds of commonsense knowledge (chapter 6). In the first of these chapters, the authors note that the 17th-century philosopher Gottfried Leibniz discovered the fundamental principles of knowledge representation: "It is obvious that if we could find characters or signs suited for expressing all our thoughts as clearly and as exactly as arithmetic expresses numbers or geometry expresses lines, we could do in all matters insofar as they are subject to reasoning all that we can do in arithmetic and geometry. For all investigations which depend on reasoning would be carried out by transposing these characters and by a species of calculus." These principles were made concrete in the development of formal logic over the following centuries but are more general than any particular system of logic. All that is required is that there are symbols that represent real-world entities and relationships over entities; rules for combining symbols into expressions that can be interpreted as representing states of affairs in the world; and rules for mechanically manipulating expressions that represent reasoning over states of affairs. In propositional or predicate logic, the symbols are strings, the expressions are sentences recursively constructed from symbols, and the rules for reasoning represent logical deduction; however, a knowledge representation system can compose symbols in other manners, such as graphs, and the manipulation rules can represent other forms of reasoning, such as prototypical (default) reasoning or probabilistic reasoning. The fact that people can talk about what they know, what other people believe, what would hold if something other held (counterfactuals), and can learn

about the real world by reading strings of symbols, can all be explained if humans employ some kind of knowledge representation and reasoning. "What alternative to the KR hypothesis is there for common sense in AI systems?', the authors write, "The answer: so far, none."

Chapter 5 outlines the general phenomena that a knowledge representation system must distinguish as presuppositions to any specific knowledge about some domain of concern; in other words, "a theory of everything". This includes that there are things, some physical ("owls, refrigerators, and lakes") and some non-physical ("numbers, beliefs, and stories"). Time points are linearly ordered nonphysical things. Things may come into or out of existence at a time point and have properties - relationships to other things - that may be fixed or may be relative to time. Events are nonphysical things that occur at a time point or over a sequence of time points that cause properties of other things to change. From this basis, the authors go on to describe what has been called "naive physics", the way we think about physical objects and their properties - even when we are doing science and engineering, except perhaps in the most extreme regions of quantum mechanics. They go on to outline naive psychology, what is commonly referred to as "theory of mind" in the cognitive science literature, and causality. The reader who has never built an AI system may be forgiven for thinking that all of this is too obvious - too necessary - to state. A computer program by itself knows none of this. In most programs, the way the world is broken down into things and relationships is highly simplified and only implicitly defined. A database system, for example, might include a table of employees and salaries, and allow a user to run a query to find the average salary of employees. This presupposes that there are things called "employees" and each has a property called "salary" which is a number. For an AI system to reason about and act in the world, however, it must internalize something like this theory of everything. Most work in the knowledge representation tradition would have the programmer - or perhaps a community of contributors - manually build the theory into the system. If we think of humans as an AI system, this would be assuming that such fundamental concepts are hardwired into our brains by evolution. Researchers in psychology or AI who favor a blank-slate notion of human or machine intelligence might say it is all learned - although as we noted above in our comment on learning disentangled representations, it is a challenge to make deep learning systems to even discover that physical objects have general properties such as size or color. The next chapter drills down into more details about what is required to represent commonsense knowledge, including general concepts (e.g., birthday parties); prototypical or default properties of such concepts (e.g., birthday parties are held on or near the honoree's birthday); and information about specific individuals or events (e.g., Sarah had her party a month after her birthday). Kinds of events include in their representation the sub-events that they typically, but do not necessarily contain (e.g., a birthday party includes games, serving cake, singing "Happy Birthday", etc.).

Chapters 7 through 9 describe a way to implement this kind of knowledge representation and a series of algorithms for performing reasoning and planning. It draws on the technical work by Brachman on semantic networks and Levesque on automated planning. The authors avoid familiar logical notation or computer languages such as LISP or Python to make the material clear to a lay reader, and in this, they succeed. The appendix summarizes the language and

algorithms, and it would be straightforward for a moderately proficient programmer to turn them into working code. Algorithms are described for bottom-up reasoning, for interpreting observations, and top-down reasoning, for generating plans.

One of the key ideas of these chapters is that commonsense reasoning enables an AI system to act appropriately in the face of unforeseen events. The authors present the example of an agent driving to a store when something unexpected occurs; after stopping at a red light, the signal does not change to green after five minutes. The agent must find an explanation for the anomaly to repair the plan and finding the explanation may require the use of many kinds of commonsense knowledge. For example, the signal light may simply be broken, so an appropriate response would be to make a U-turn, go down a block, and then cross at the next intersection. Alternatively, suppose the driver hears an approaching marching band and the date is July 4th. In this case, the best explanation for the failure is that the street has been closed for a parade, and the agent should find a route to the store that does not involve crossing it, or if one does not exist change its goal to go to a different but similar store.

While the examples of commonsense reasoning in action in these chapters are thought-provoking, many people (including myself) may find the proposed reasoning algorithms inadequate. There are an unlimited number of explanations for any observation or failure of expectations. It is surely necessary to have a general and consistent way of choosing the most likely interpretation or explanation - in other words, to reason probabilistically. There is a long but unsuccessful history in AI of attempts to use prototypes and defaults to handle uncertainty. These attempts failed because absent quantitative probabilities, one must include explicit rules that break ties between every competing hypothesis under every possible condition of the world. The number of prototypical or default rules becomes astronomical. Between the height of the knowledge representation and reasoning era in AI and the current deep learning era was what was known as the Bayesian revolution in AI, when work on Bayesian probabilistic methods for reasoning and learning dominated the field [Pearl 1995]. The Bayesian revolution was sparked by Judea Pearl's introduction of Bayesian networks, a way to represent knowledge about probability in a graph-based form. Bayesian networks and related formalisms extend classical probability theory with what is called the maximum entropy assumption, which is a kind of single, general default rule that formalizes the notion of "all things being equal". Bayesian networks are a perfectly good knowledge representation in Leibnitz's sense and have the advantage of providing a probabilistic sound way to use the context of a knowledge base - what is not known as well as what is known - for interpretation or planning. (Prototypes or defaults can be thought of as a coarser instrument for using such context.) Bayesian networks are a subclass of a more general class of representations named graphical models, so-called because their syntax is based on graphs rather than on sentences as in propositional or first-order logic. One of the hot topics in research in deep learning is determining the conditions under which artificial neural networks can be viewed as implementing inference in a graphical model; success in this endeavor would allow us to understand neural networks as a knowledge representation scheme, rather than as an algorithmic black box [e.g., Patel 2016].

It is unfair, however, to criticize a book for what it does not try to do. Machines Like Us succeeds brilliantly in making the results of decades of work in knowledge representation accessible and relevant to lay readers and the current generation of AI researchers. Almost no work in deep learning has even attempted to develop a "theory of everything" for the physical, mental, or social worlds, as the authors have attempted in this book. I hope it may inspire others - in particular ones equipped with the tools of probabilistic models and machine learning - to take on the challenges.

References

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS 2012), December 2012, pages 1097–1105.

David Silver, Julian Schrittwieser, K. Simonyan, Ioannis Antonoglou, Aja Huang, A. Guez, T. Hubert, Lucas baker, Matthew Lai, A. Bolton, Yutian Chen, T. Lillicrap, Fan Hui, L. Sifre, George van den Driessche, T. Graepel, D. Hassabis. Mastering the game of Go without human knowledge. Nature, Volume 550, 2017, pages 354-359.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35(8):1798–1828, 2013.

Douglas Hofstadter. Artificial neural networks today are not conscious, according to Douglas Hofstadter. The Economist, June 9, 2022.

Ronald J. Brachman and Hector J. Levesque. Machines Like Us: Toward AI with Common Sense. MIT Press, 2022.

Gary Marcus and Ernest Davis. Rebooting AI: Building Artificial Intelligence We Can Trust. Vintage, 2019.

Ronald J. Brachman and James G.Schmolze. An overview of the KL-ONE Knowledge Representation System. Cognitive Science, Volume 9, Issue 2, April–June 1985, pages 171-216,

H. Levesque, R. Reiter, Y. Lespérance, Fangzhen Lin, and R. Scherl. GOLOG: A Logic Programming Language for Dynamic Domains. Journal of Logic Programming, Volume 31, 1997, pages 59-83.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The Winograd schema challenge. KR'12: Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning, June 2012, pages 552–561.

John McCarthy. Programs with Common Sense. Presented at the Symposium on the Mechanism of Thought Processes, National Physical Library, 1958.

Judea Pearl. Probabilistic reasoning in intelligent systems: Networks of plausible inference. Synthese-Dordrecht, 1995.

Ankit B. Patel, Tan Nguyen, and Richard G. Baraniuk. A Probabilistic Framework for Deep Learning. Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016), December 2016.