The third AI summer: AAAI Robert S. Engelmore Memorial Lecture

Henry A. Kautz

Department of Computer Science, University of Rochester, Rochester, NY, USA

Correspondence
Henry A. Kautz, Department of Computer Science, University of Rochester, Rochester, NY 14627, USA.
Email: henry.kautz@gmail.com

INTRODUCTION

This article summarizes the author's Robert S. Englemore Memorial Lecture presented at the Thirty-Fourth AAAI Conference on Artificial Intelligence on February 10, 2020. It explores recurring themes in the history of AI, real and imagined dangers from AI, and the future of the field.

We are now in AI's third summer, a period of rapid scientific advances, broad commercialization, and exuberance—perhaps irrational exuberance—about our potential to unlock the secrets of general intelligence. Twice before the field of AI has experienced such a period, and each was followed by a winter of collapse of commercialization and drastic cuts in government investments in research. In this essay, I will argue that despite this cyclical history, enduring insights have blossomed each summer. The winters can be viewed as times of contemplation and integration that advance through the synthesis of new and old ideas. I will also argue that we may be at the end of the cyclical pattern; although progress and exuberance will likely slow, there are both scientific and practical reasons to think a third winter is unlikely to occur.

In every summer, articles and books about AI written for nonexperts have found wide audiences. I read four recent books shortly before writing this essay: *The Master Algorithm*, by Domingos (2015); *AI Superpowers*, by Lee (2018); *Human Compatible*, by Russell (2019); and *Rebooting AI*, by Marcus and Davis (2019). This first is an objective history of machine learning, and like this essay, emphasizes the continuous evolution of the field. The second charts the dramatic rise in AI R&D in China and points the way

to a utopian future. The third argues that superhuman artificial intelligence will be an existential risk if the values of such AIs are not aligned by design with those of humans. The fourth contends that deep learning, the most powerful approach to machine learning devised to date, will soon reach inherent limits, and that a different approach that synthesizes recent and older approaches to AI will be necessary in the future. This essay will touch on many of the same elements as these three books. I will first provide a history of AI; next, discuss near-term dangers of AI; and finally, describe a number of different technical approaches for future AI.

If one was to create a cartoon history of AI, the first panel would show the symbolic approach to AI—pictured, say, as a cat, beating up on the artificial neural network approach—let us picture it as Jerry the Mouse. The second panel shows both Tom and Jerry shivering in a wintery scene; and the third shows Jerry, now grown huge and powerful through deep learning, easily dispatching Tom (Figure 1). There is more than a grain of truth in this cartoonish view of the history of AI from the 1980s through the present day. The story it presents is incomplete, however, both in chronology and in failing to illustrate the rich set of ideas and approaches that developed and entwined through the history of the field.
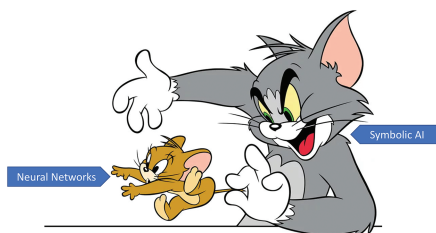
THE FIRST AI SUMMER: IRRATIONAL EXUBERANCE, 1948–1966

William Grey Walter was a polymath in neuroscience and electronics. As a young man in the 1930s, he built the first

(A) The Cartoon History of AI
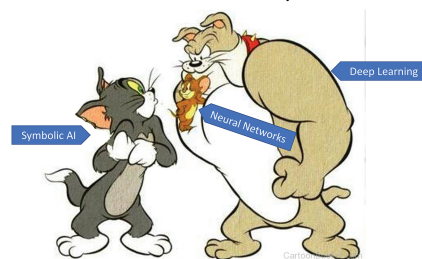

(B) The Cartoon History of AI

FIGURE 1 Cartoon history of AI picture: (a) Tom (symbolic AI) beating Jerry (neural networks). (b) Spike (deep learning) holding Jerry and beating Tom

FIGURE 2 Picture of William Grey Walter and his tortoise

electroencephalography (EEG) in the United Kingdom and discovered that the measurement of brain waves could be used to locate brain tumors responsible for epilepsy (Walter 1953). Thirty years later, a groundbreaking paper he coauthored in *Nature* showed that spikes in neural activity could be used to predictive motor events a full half-second before the subject was consciously aware of having made the decision to move—in other words, that the conscious mind only *thought* it was making decisions (Walter et al. 1964).

Walter was as much an engineer and tinkerer as a scientist. During WW II, he designed radar systems. The mechanistic view he took of the brain led him to experiment with artificial neural networks—not just as a mathematical abstraction, as in the work of McCulloch and Pitts (1943), but as the decision-making engine for an embodied artificial animal (Figure 2). Beginning in 1948, he built and demonstrated a series of increasingly sophisticated autonomous tortoise-shaped three-wheeled robots (Hoggett 2011). Their analog electronic brains employed up to seven vacuum tubes, which interpreted signals from touch, light, and sound sensors and controlled propulsion

and steering motors. Although their behavior was hard-wired, the later versions supported a form of conditioned-reflex learning. A capacitor-based memory could learn to associate the simultaneous activation of two sensors—for example, the sound of a whistle and the obstacle detecting bump sensor. The reflex triggered by the bump sensor—backing up and turning—could then be triggered by the sound sensor.

Artificial neural networks

The tortoises' legacy includes the field of artificial neural networks, which today dominates research and development in artificial intelligence. A few of the well-known major steps in the development of artificial neural networks were the error-based perceptron learning rule of Rosenblatt (1958), the development of backpropagation for training multilayer networks (Werbos and Werbos 1974; Rumelhart, Hinton, and Williams 1986), and parameter sharing in structured networks, and in particular, convolutional networks (Fukushima 1980; LeCun et al. 1989). It is easy to see the aspects of artificial neural networks that Walter got wrong: most obviously, the use of analog electronics and the focus on stimulus—response learning rather than error-minimization learning. It is just possible, however, that Walter was simply wildly premature. Artificial neural networks are now being compiled into edge-computing hardware for applications such as video surveillance; while such hardware is now digital, there is research on creating analog artificial neural networks that could operate with a fraction of the energy needed by digital circuits. Furthermore, one could argue that the tortoises' implementation of stimulus–response learning was an early attempt at unsupervised learning—which is today the most important and challenging problem in research on machine learning.

Artificial neural networks were not the only legacy of the tortoises. They demonstrated that complex purposeful behavior arises in the interaction between an agent and

an environment, an idea that stands in sharp contrast to the more cerebral symbolic approaches to AI that we will describe shortly. Walter was part of a larger movement that aimed to understand animal and machine intelligence using feedback loops and other tools of control theory. The field was given the name "cybernetics" with the publication of Norbert Weiner's book of that name (Weiner 1948). The tortoises were a perfect example of a mechanism regulated by feedback from their environment. Cybernetics flourished in the former Soviet Union, but never gained a foothold in the US AI research community until a synthesis of control theory and dynamic programming (Bellman 1957) emerged under the banner of reinforcement learning (Witten 1977; Sutton and Barto 1981). Even then, researchers in reinforcement learning were a small minority in the general AI community for decades. Researchers made steady progress in developing mathematical frameworks for training control systems when the feedback signal was distant in the future. The fact that rewards can be temporally distant from the agent's actions distinguishes reinforcement learning from stimulus–response learning; indeed, the ability to act for delayed gratification is a key aspect of intelligence. Temporal-difference learning (Sutton 1988) provided a general approach for implementing gratification, and proved to be particularly effective when the agent's internal state was represented by a neural network. We shall see the potent combination of artificial neural networks and reinforcement learning reemerge in the third AI summer.

Knowledge representation

The first AI summer also saw the birth of a very different approach to building intelligent machines, an approach whose heritage stretched back thousands of years. This is the logic-based approach to AI, or more generally and accurately, the approach based on declarative knowledge representation.

Symbolic logic grew out of the art of rhetoric in ancient Greece. Around 350 BC, Aristotle formalized certain kinds of deductive arguments symbolically in his *Prior Analytics*. His key insight—indeed, the insight that is the basis for not only logic but for the theory of computing—is that reasoning could be performed by considering only the syntactic form of statements without considering the meaning of those statements. After this prescient beginning, however, over 2000 years passed before significant advances were made in formal logic. George Boole created a complete characterization of proposition logic in 1845 (Boole 1854), as Gottlob Frege, Charles Sanders Peirce, David Hilbert, and others did for quantified logics in the following decades. This generation of philosophers, however, had a primary motivation for their work that differed from that of Aristotle and his medieval followers: their ultimate goal was to provide a complete and rigorous basis for mathematics rather than to understand everyday reasoning and argumentation. They, therefore, poured enormous energy into trying to overcome the paradoxes of naive set theory (Russell 1903) and were devastated by the discovery that no logic could capture all mathematical truths (Gödel 1931).

The concerns of the researchers who pioneered the logical approach to AI stood in sharp contrast to those of the philosophers of mathematics. First, the AI researchers were encouraged by the creation of programs that could automatically find proofs of some—not necessarily all—mathematical theorems, and were untroubled by logic's inherent incompleteness. The celebrated Logic Theorist program (Newell and Simon 1956) was able to prove 38 elementary theorems from Principia Mathematica (Whitehead and Russell 1910–1913). Second, most AI researchers had little interest in mathematics as a subject matter of logic. Instead of trying to axiomatize abstruse mathematics, John McCarthy argued, researchers should strive to develop logical representations of commonsense knowledge (McCarthy 1958). McCarthy's original paper described knowledge about locations (e.g., one can be at a desk, in a car, etc.) and physical movement (e.g., one might walk from one location to another nearby location), and his former student Patrick Hayes called for axiomatization of commonsense physics (Hayes 1978). Others attempted to represent the logical rules of human discourse (Allen et al. 1977), thus closing the loop with the ancient Greeks' view of logic as a tool for analyzing rhetoric.

Researchers in the first AI summer also began work on systems that employed graphs rather than the strings or trees of classical logic to represent knowledge. These new kinds of representations were called "semantic networks" and used vertices to represent concepts and edges to represent relationships. The word "semantic" came from their initial use as an intralingua for translating between different natural languages (Richens 1956); they were intended to capture the meaning, or semantics, of sentences. Although they were presumably unaware of it at the time, one researcher has argued that semantic networks were a rediscovery of the diagrams that ancient Sanskrit scholars used to analyze texts (Brigs 1985). Researchers increasingly converged on the view that semantic networks were simply an alternative notation for classical logic, as exemplified by Ronald Brachman's work on KL-ONE (Brachman and Schmolze 1985). Just as all practical programming languages were Turing-complete and thus theoretically equivalent but differed in ease or naturalness of use, these researchers argued that semantic networks were simply a more natural form of first-order

logic where syntax explicitly described concepts in terms of their attributes and how they generalized or specialized other concepts. This version of semantic networks became known as "description logic." Pure description logic, however, proved inadequate for representing large real-world domains because it could only capture the absolutely necessary properties of concepts, not those that were prototypical or that held by default. In recent years, companies including Google, Facebook, Microsoft Bing, eBay, and IBM have developed enormous networks called "knowledge graphs" which they use to drive many applications, such as web search and product recommendation (Singhal 2012). Despite their scale and ubiquity of use, many aspects of knowledge graphs remain informal; for example, in addition to the issue of whether links represent absolute or prototypical relations, the distance between concepts in a knowledge graph is often used as a heuristic measure of concept similarity. Later in this essay, we will describe a different family of graph-based knowledge representation formalisms called "graphical models" that combine logic, graph theory, and probability theory.

Heuristic search

The first AI summer's third research campaign was the quest for efficient algorithms for combinatorial search. We now understand that in terms of formal computational complexity theory, the quest is an impossible dream: the general task of reasoning in any suitably expressive formal system is NP-complete or harder (Cook 1971), and thus, it is believed, requires worst-case exponential time. Even the problem of STRIPS-style planning—that is, finding sequences of actions that are defined in terms of preconditions and effects—for the simple "blocks world" domain is NP-complete (Gupta and Nau 1991). However, the fact that such complexity results had not yet been discovered may have helped lead the early AI researchers to an important insight: an enormous space of possibilities could be searched in ways that were more efficient than simple enumeration. This insight differentiated AI researchers from philosophers and mathematicians, for whom the existence or nonexistence of an algorithm that would terminate after an exhaustive enumeration of possibilities was the end of the discussion: this problem was decidable, or that problem was undecidable. It was obvious to AI researchers that human reasoning was not a simple enumeration, but involved shortcuts that made the task feasible given the time and computational resource limitations of the brain. Herbert Simon, J. C. Shaw, and Alan Newell discovered and implemented one such search algorithm, means-ends analysis, in their General Problem Solver (1959), and later Newell and Simon argued that humans employed it as well

FIGURE 3 Shakey the robot implemented the STRIPS planning algorithm (Fikes and Nilsson 1971)

as a variety of other reasoning strategies in their monumental treatise Human Problem Solving (1972).

How can non-enumerative search be practical when the underlying problem is exponentially hard? The approach advocated by Simon and Newell is to employ heuristics: fast algorithms that may fail on some inputs or output suboptimal solutions. For example, the means-end planning heuristic chooses an action, which will reduce the difference between the initial and goal state; applies the action initial state; and recursively applied the process to the new state and the goal state (Figure 3). Although intuitively appealing, it is not difficult to find problems where the heuristic fails, stuck in a cycle where it reduces one difference but introduces another. The A* algorithm (Hart, Nilsson, and Raphael 1968) provided a general frame for complete and optimal heuristically guided search. A* is used as a subroutine within practically every AI algorithm today but is still no magic bullet; its guarantee of completeness is bought at the cost of worst-case exponential time.

An interesting class of incomplete heuristic search algorithms is those based on a "noisy" version of iterative repair, a heuristic similar to means-end analysis. Iterative repair begins by guessing a solution to the problem. It then iteratively identifies a flaw in the solution and patches it, yielding a new proposed solution. As with means-end analysis, simple iterative repair can easily become stuck in a cycle. A noisy version of iterative repair reduces the likelihood of becoming stuck by periodically making random changes in the solution; even if most of the random changes are bad, eventually a change is likely to be introduced that lets the search break out of the cycle. A version of noisy iterative repair named "simulated annealing" was invented by physicists and has proven widely applicable for optimization problems (Kirkpatrick, Gelatt,

and Vecchi 1983). Bart Selman and I showed that a simple version of iterative repair called "local search with noise" was even more effective for finding satisfying assignments to logical formulas; the reason for the improvement was that the random steps were restricted to ones that made the proposed solution satisfy at least one previously unsatisfied problem constraint, even if did the reverse for some of the other constraints (Selman, Levesque, and Mitchell 1992; Selman, Kautz, and Cohen 1996[1]).

Another way to solve in practice an NP-hard problem is to employ an algorithm whose empirical complexity on a problem distribution of interest grows subexponentially or exponentially with a very small exponent. For example, the best complete algorithm for satisfiability testing is backtracking search over the space of partial variable assignments, the Davis–Putnam–Logemann–Loveland algorithm (DPLL) (Davis et al. 1961), augmented by a technique called "clause learning" (Marques-Silva and Sakallah 1996; Bayardo and Schrag 1997). When the backtracking algorithm reaches a dead end—that is, when it determine that the current partial assignment is inconsistent - the clause learning module computes a minimal subset of previous assignment choices that led to the inconsistency and adds the negation of that combination to the problem as a new clausal constraint. The new clause prevents those choices from being made in a different branch of the search tree, thus pruning the search. Although still an exponential algorithm, my colleague Paul Beame, student Ashish Sabharwal, and I showed that DPLL with clause learning is probably more powerful than DPLL (Beam, Kautz, and Sabharwal 2004). The algorithm demonstrates remarkably restrained growth in many real-world problem domains. For example, Pushak and Hoos (2020) argued that the algorithm on bounded model-checking problems shows subexponential empirical scaling.

THE FIRST AI WINTER: CRUSHED DREAMS, 1967–1977

During the first AI summer, many people thought that machine intelligence could be achieved in just a few years. The Defense Advance Research Projects Agency (DARPA) launched programs to support AI research with the goal of using AI to solve problems of national security; in particular, to automate the translation of Russian to English for intelligence operations and to create autonomous tanks for the battlefield. Researchers had begun to realize that achieving AI was going to be much harder than was supposed a decade earlier, but a combination of hubris and disingenuousness led many university and think-tank researchers to accept funding with promises of deliverables

that they should have known they could not fulfill. By the mid-1960s neither useful natural language translation systems nor autonomous tanks had been created, and a dramatic backlash set in. New DARPA leadership canceled existing AI funding programs. In 1969, the powerful Senate Majority Leader Mike Mansfield hobbled AI research funding by all military agencies for decades by pushing through a law that prohibited military funding of fundamental research beyond specific military functions.

Outside of the United States, the most fertile ground for AI research was the United Kingdom. The AI winter in the United Kingdom was spurred on not so much by disappointed military leaders as by rival academics who viewed AI researchers as charlatans and a drain on research funding. A professor of applied mathematics, Sir James Lighthill, was commissioned by Parliament to evaluate the state of AI research in the nation. The report stated that all of the problems being worked on in AI would be better handled by researchers from other disciplines—such as applied mathematics (Lighthill 1973). The report also claimed that AI successes on toy problems could never scale to real-world applications due to combinatorial explosion. This claim, of course, ignored the quest in AI for methods to tame combinatorial search as described above. In response to the report, all public funding of AI research in the United Kingdom was terminated.

SECOND SUMMER: KNOWLEDGE IS POWER, 19[67]8–1987

The second AI summer was marked by the field's change in focus from commonsense knowledge to expert knowledge. Expert systems, it was believed, would be able to substitute for trained professionals in medicine, finance, engineering, and many other fields. An expert—say, a doctor—would be debriefed by a knowledge engineer, who would encode the expert's vast experience into a large set of rules and facts. A general symbolic reasoning system could then apply these rules to solve particular problems—for example, to create a diagnosis on the basis of a patient's symptoms. The rules could also drive the system to gather further information—for example, to order certain blood tests for the patient in order to refine the diagnosis.

The date for the beginning of the second summer is written in the section heading using regular-expression notation to mean that it could be said to have started in 1968 or to have started in 1978. In 1968, Feigenbaum, Lederber, and Buchanan (1968) created the first expert system, Dendral. It was intended to help organic chemists in identifying unknown organic molecules by analyzing their mass spectra and using knowledge of chemistry. Dendral gained much academic interest and led to the

development of expert systems in other domains, notably MYCIN (Shortliffe and Buchanan 1975) for bacterial infection diagnosis and INTERNIST-I, which aimed to capture the internal medicine expertise of the chair of the department of internal medicine at the University of Pittsburgh (Pople 1976).

It was not until 1978, however, that expert systems became a hot area of R&D with the creation and commercial deployment of XCON (McDermott 1980). In the 1970s, buying a computer system was a slow and error-prone process. Computers were much less standardized than they are today, and a buyer needed to choose among hundreds of options when placing an order for one. Options could interact in complex ways: some combinations of options could not be physically built or if built would lead to poor performance; some options required choices from other options; and so on. The process to order a VAX computer from Digital Equipment Corporation (DEC) could require as long as 90 days of back-and-forth between a customer, sales representatives, and DEC engineers to create a correct system configuration. XCON reduced the time to generate a satisfactory system configuration for a customer to about 90 minutes. The enormous advantage this gave DEC in the marketplace did not go unnoticed. Soon, companies of all sorts began developing and deploying expert systems for a variety of tasks in engineering and sales. Feigenbaum's phrase "knowledge is power" became the slogan of the era.

The second AI summer differed from the first in that it was driven as much by commercial money as by government support. In addition to investments by companies using expert systems, venture capital flowed into companies creating a software and hardware ecosystem to support expert systems. Software startups sold expert system "shells," that is, reasoning engines with user interfaces intended to make it possible for nonprogrammers to enter rules. The fact that expert system development was incremental meant that dynamically linked programming languages were preferred—which in the 1970s and 1980s meant varieties of LISP or Prolog. The relatively slow performance of these languages with the implementations and hardware of the era motivated building computer hardware to directly interpret LISP (by the startup companies Symbolics and LMI) or Prolog (by various Japanese companies under the auspices of Japan's Fifth Generation project).

THE SECOND AI WINTER: THE DISRESPECTED SCIENCE, 1988–2011

Many reasons can be offered for the arrival of the second AI winter. The hardware companies failed when much more cost-effective general Unix workstations from Sun

together with good compilers for LISP and Prolog came onto the market. Many commercial deployments of expert systems were discontinued when they proved too costly to maintain. Medical expert systems never caught on for several reasons: the difficulty in keeping them up to date; the challenge for medical professionals to learn how to use a bewildering variety of different expert systems for different medical conditions; and perhaps most crucially, the reluctance of doctors to trust a computer-made diagnosis over their gut instinct, even for specific domains where the expert systems could outperform an average doctor. Venture capital money deserted AI practically overnight. The world AI conference IJCAI hosted an enormous and lavish trade show and thousands of nonacademic attendees in 1987 in Vancouver; the main AI conference the following year, AAAI 1988 in St. Paul, was a small and strictly academic affair.

Commercial factors aside, enthusiasm for expert systems cooled because of two central technical challenges; indeed, overcoming these challenges set the workplan for the next two decades of research in AI. The first challenge was the need for principled and practical methods for probabilistic reasoning. The logical rule-based approach excelled at capturing knowledge about relationships among concepts and entities (such as class/subclass/instance or object/part/attribute hierarchies) but was poorly suited for problems where one needed to assign probabilities to conclusions. Although the need to handle uncertainty was recognized by early expert system researchers, they did not yet know of probabilistically sound methods of reasoning that were computationally practical; systems such as MYCIN and its descendents instead attached "certainty factor" numbers to rules and facts and combined them in an ad hoc manner. The second unsolved challenge for the expert system approach was named the "knowledge acquisition bottleneck." Capturing all but the narrowest domains required a huge number of rules. Not only was it difficult or impossible to recruit and train enough experts to write enough rules, but once the knowledge bases became large they inevitably became full of inconsistencies and errors.

Probabilistic reasoning

The field of AI did not disappear during the slightly more than two decades of the second AI winter. It continued steadily as a relatively small but intellectually vigorous research field, freed of the hype and demands for commercial profit. The challenge of sound but efficient probabilistic reasoning was first met by what were called graphical probabilistic models. Bayesian Networks (Pearl 1988) provided a solution to the problem of compactly

representing multivariable probability distributions without requiring exponentially large probability tables. Each conditional probability statement was represented by a set of directed edges that ended at a node together with a conditional probability table for the variable associated with the node. The graph has a much stronger meaning, however, than the conditional probability statements alone: it represents a single probability distribution—the so-called maximum-entropy distribution—rather than all distributions that are consistent with the original conditional probability statements. In many problems, one does indeed want to reason with a maximum-entropy distribution because it is the one under which our given knowledge captures all interesting relationships between the variables. The introduction of Bayesian networks led to fruitful decades of research on extensions to Bayesian networks, alternative graphical models, and a variety of new algorithms for probabilistic reasoning. Heckerman and Shortliffe (1992) discovered the conditions under which MYCIN's certainty factors could be given a probabilistic interpretation, thus explaining why expert systems sometimes gave sensible answers but at other times did not. Around the turn of the century, the field of statistical–relational reasoning arose, which sought to develop representations and algorithms that combined the semantics of graphical models with the expressive ability of the finite fragment of first-order logic (Friedman et al. 1999; Richardson and Domingos 2006).

Machine learning

Overcoming the knowledge acquisition bottleneck led the field of AI to a renewed focus on machine learning. For most of the second winter, however, few researchers returned to the roots of machine learning in artificial neural networks. Methods were developed for learning decision trees (Quinlan 1986) and logical rules (Muggleton and Feng 1990). The parameters (conditional probability tables) for graphical models could be directly estimated from complete data or estimated by the expectation–maximization algorithm for incomplete data (Dempster, Laird, and Rubin 1977). Valiant's (1984) work on probably approximately correct (PAC) learning showed the limits of learnability for any method relative to the amount of data that were available. Until the revival of artificial neural networks in the third summer, the most powerful approach to "black box" machine learning, that is, that did not rely upon or attempt to create an interpretable domain model, was the support vector machine (SVM) pioneered by Cortes and Vapnik (2004). An unintuitive feature of SVMs is that they often worked well when highly over-

parameterized—a situation that had been thought to be necessarily associated with overfitting. Deep learning with artificial neural networks turned out to share this surprising feature.

Even as AI research methodology became steadily more rigorous and many powerful new methods for learning and reasoning were developed during the second AI winter, distain of the phrase "artificial intelligence" remained strong in the commercial world. When AI began to make the leap to large-scale, real-world applications, companies often went to pains to promote such systems using terms other than AI. For example, when IBM wished to leverage the success of its Watson general question–answering system (Ferrucci et al. 2013), it invented the phrase "cognitive computing" and used it exclusively instead of AI until quite recently.

THE THIRD SUMMER: DEEP LEARNING (201[26] - ?)

As is the case with the Second AI Summer, the start of the Third Summer can be dated either to when the initial technical breakthrough occurred or the date when the rest of the world took note. The technical breakthrough occurred in 2012. A few years earlier, the computer vision community had created the ImageNet challenge, a benchmark for classifying more than a million images of single objects across 1000 categories (Russakovsky et al. 2015). A distributed group of researchers including had begun experimenting with running many-layered artificial neural networks on graphic processor cards, and entered the competition in that year. The ANN model, AlexNet (Krizhevsky, Sutskever, and Hinton 2012) demolished the competition of traditional computer vision algorithms, achieving an error rate of 16% versus 26% by the best competitor. Within months, most of the computer vision research community was working on many-layered ANNs, and thus the field of what is now called deep learning ("deep" referring to the many layers of the models) was born. The year the world at large took AI seriously again was 2016. That was the year that Google DeepMind AlphaGo defeated the Go grandmaster Lee Sedol (Figure 4). AlphaGo performed a stochastic variation of game tree search with a deep neural network to evaluate leaf positions, where the deep ANN was trained through self-play (Silver et al. 2016). It would be hard to overstate the impact that AlphaGo's victory made in China, or how widely China's excitement and subsequent enormous investments in AI reverberated in governments and companies throughout the world. The AlphaGo match has been called China's "Sputnik Moment" (Lee 2018).

FIGURE 4 Picture from the Lee Sedol versus AlphaGo match. Credit: Google DeepMind

The Third Summer reignited the First Summer themes of artificial neural networks and reinforcement learning. What other new and lasting insights did it ignite?

Hierarchical representation learning

Many would say the most important advance deep learning provides is hierarchical representation learning. While logical knowledge representation was all about capturing hierarchies, earlier approaches to machine learning considered only two levels of representation: feature and class. It was the job of the machine learning scientist to engineer features that were at the most useful level of abstraction (Figure 5). The primary motivation for deep learning was to eliminate the need for manually engineered features; indeed, one of the premiere conferences on deep learning, founded by Yann LeCun, is named the "International Conference on Representation Learning."

Similarity

A second facet of deep learning is less appreciated but no less a breakthrough for AI: the fact that deep learning representations directly support concept similarity. Since everything is a vector of activations, the distance between vectors—computed by the vector cosine or other operation—is easily computable. Reasoning about similarity is vital for many (arguably most) real-world domains. A toy example showing vectors for "cat," "kitten," "dog," and "house" appears in Figure 6. If the task is to choose a
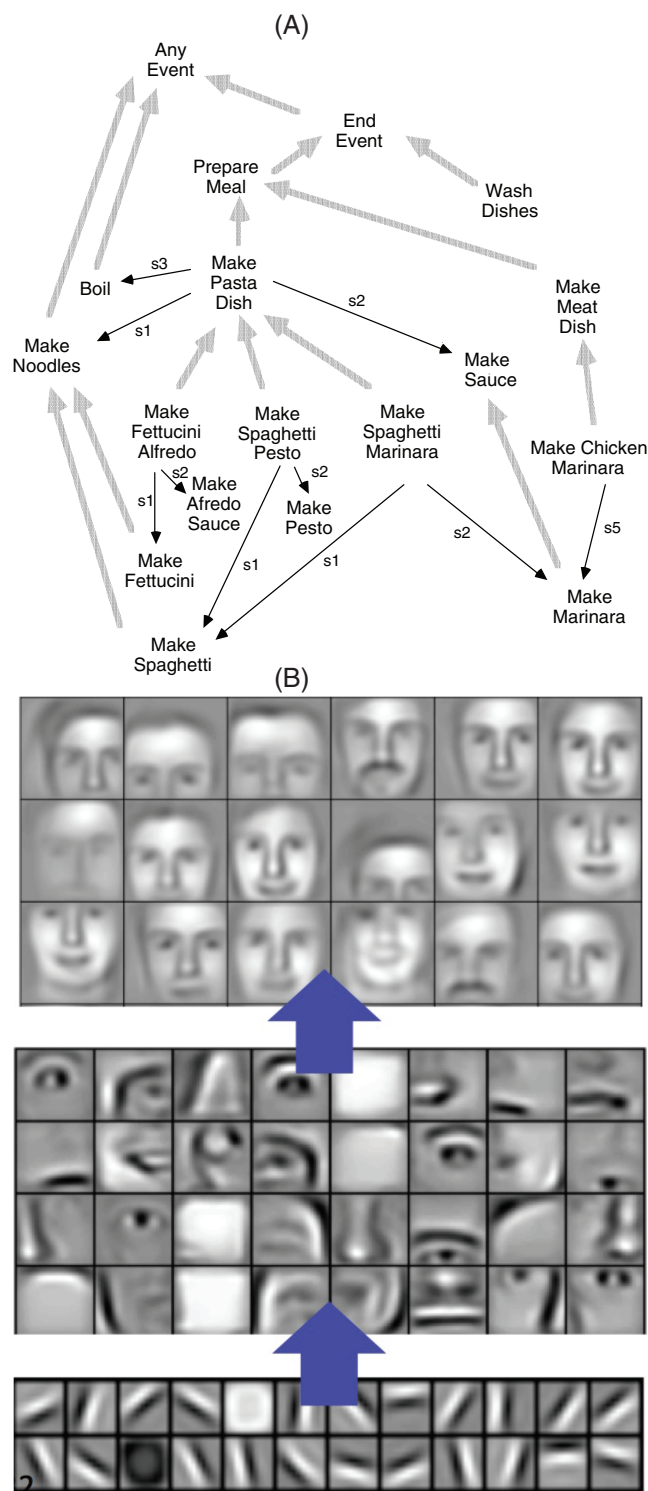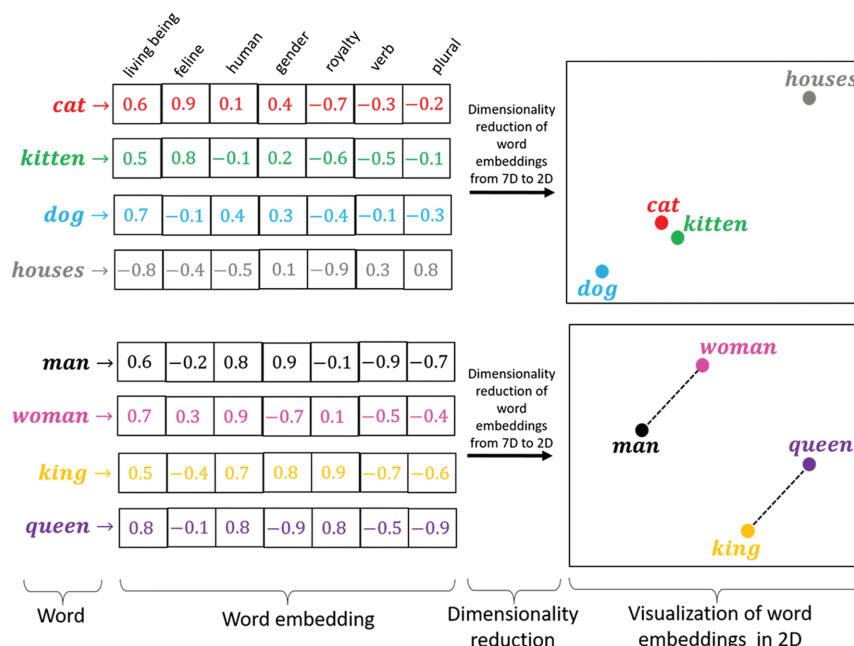

FIGURE 5 (a) Diagram of logic-based action hierarchy containing subclass and substep links from Kautz (1987). (b) Illustration of representation learning from Lee et al. (2011)

gift for your daughter and she asks for a kitten, then giving her a cat instead is a reasonable substitute; giving her a dog is riskier; and giving her a house is not a good idea at all. More seriously, learning can be viewed as generalizing from known cases to novel but similar cases.

FIGURE 6 Illustration of similarity

	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

Word | Word embedding | Dimensionality reduction | Visualization of word embeddings in 2D

Similarity is not captured by probability. A dog does not have some percent chance of being a cat. The only prior work in AI that had seriously dealt with similarity as distinct from probability was Lofti Zadah's Fuzzy Set Theory (Zadeh 1965).

WHY WINTER MIGHT NOT RETURN

Are we on the verge of another AI Winter? Although we are still far from what has been called "general artificial intelligence" (AI systems that have the power and flexibility of the human brain) the kind of AI that today's technology can deliver is good enough for solving a huge number of practical problems ranging from natural language translation to managing investment portfolios. Current artificial neural net systems are dramatically different from brains in scale, organization, and the algorithms they implement for learning and inference. ANNs also differ from brains in that they are not generally embodied in a physical organism, and if neuroscience and psychology have taught us anything in recent decades, it is that the human mind is not separate from the human body. How then is it possible for ANNs to imitate certain aspects of human intelligence so well?

The first answer I would offer to this puzzle is that ANNs may be capturing general principles for intelligence that are independent of the particular structure of the human brain. An analogy may be made with the evolution of intelligence in the octopus, whose common ancestor with humans was a worm-like creature that lived 300 million years ago and had only a rudimentary nervous system (Vitti 2013). The octopus has a ring-shaped brain with a lobe at the base of each arm, and lives a life almost totally alien to that of humans or mammals in general. They are born in groups of around 50,000 and are not cared for by their parents; they grow to maturity in about 2 years, and except for a brief moment of mating, live solitary lives until dying of old age at around 5 years. Yet, through a seeming miracle of convergent evolution, they demonstrate the hallmarks of intelligent behavior, including learning, planning, and tool use, and according to some people who have spent much time with them, they can come to recognize particular humans and interact with them in the playful manner of one intelligent creature to another (Ehrlich and Reed 2020). ANNs, like the octopus, may be an instance of technology-driven convergent evolution of intelligence.

The second answer I would offer is that true intelligence is often unnecessary because of the unreasonable predictability of the world. The Transformer models of natural language are essentially statistical models of word co-occurrence (Devlin et al. 2019). Given a sequence of words, they estimate the probability distribution over the next word to occur; they can be applied iteratively to generate texts by randomly picking the next words according to this distribution. Who would have dreamed that such statistical monkeys at typewriters could create long coherent passages of English as in the famous example from OpenAI's GPT-2 language model (Radford et al. 2019).

Prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect

English. *GPT-2*: While examining these bizarre creatures, the scientists discovered that the creatures also spoke some fairly regular English. Pérez stated, "We can see, for example, that they have a common language, something like a dialect or dialectic." Dr. Pérez believes that the unicorns may have originated in Argentina, where the animals were believed to be descendants of a lost race of people who lived there before the arrival of humans in those parts of South America.

Whether my first or second answer is correct—or some combination or neither!—the fact remains that artificial intelligence is now good enough to solve practical problems in a wide variety of domains. At worse, the hype around AI may diminish, but research and commercial support will not diminish. A reduction in hype would in fact be a good force for the field. We might think of it as a coming time of pleasant summer weather, a relief after the current heat wave.

AI FOR BAD

Kai-Fu Lee's book, AI Super-Powers, makes a convincing case that AI will transform practically every aspect of life, work, and human relationships. He ends on an optimistic note, envisioning a future where AI handles all the drudgery of work, and humans are all employed in meaningful jobs that involve close and warm human interaction, such as teachers, caregivers, and artists. Such an optimistic view of the impact of AI is quite widespread in the AI community. There are a growing number of scientific workshops and conference tracks on the theme of "AI for Good," and the phrase is also the name of an annual United Nation Global Summit, the name of a Microsoft initiative, and appears in the mission statement of OpenAI. Rather than continuing in such a vein, let us instead turn to "AI for Bad." In this section, I shall argue that we do indeed face terrible threats from AI, but these threats are not the ones most widely discussed in either popular or academic writing.

Keeping your (face | mind) private

In 2018, Joy Buolamwini and Timnit Gebru observed that several experimental and commercial face recognition systems have lower accuracy on darker-skinned faces than lighter-skinned faces, a finding that was supported in part by a National Institute of Standards and Technology evaluation of face recognition systems the following year (Grother, Ngan, and Hanaoka 2019). This led to popular outcry against face recognition technology under the supposition that police would use face recognition to arrest and convict innocent darker-skinned people (Chinoy 2019, Crockford 2020). Laws were passed in various municipalites, such as San Francisco (Conger, Fausset, and Kovaleski 2019), against the use of face recognition software by police and other agencies. The fact that a trained ML system's accuracy across subgroups reflects the relative size of those subgroups in the training data came as no surprise to experienced researchers. For example, face recognition systems developed and trained on datasets in China are more accurate in recognizing Asian faces than systems built in the United States. Further, much of the reporting of the issue oversimplified the concept of accuracy. In any recognition system, there is a tradeoff between the false-positive rate and the false-negative rate that depends upon a threshold parameter. The NIST study actually showed that depending upon the choice of threshold, the false match rate for Black faces could be higher or lower than for White faces (Grother, Ngan, and Hanaoka 2019, Annex 12: Error tradeoff characteristics with US mugshots, Figure 1). The ban on face recognition technology in law enforcement is particularly tragic in light of the fact that mistaken human eye-witness identification is known to be very high (Albright 2017). To date, there has been one reported case of an innocent man being arrested due to an error by face recognition software (along with the human error in confirming the results of the identification); the victim spent several hours in a jail cell before being released (Hill 2020). By comparison, the Innocence Project has found that mistaken eyewitness identifications contributed to about 70% of the wrongful convictions in the United States that were overturned by postconviction DNA evidence (Innocence Project 2020).

While face recognition is an AI threat that has been much exaggerated, there is a related real and present AI-driven threat that cannot be overemphasized. Instead of concern about keeping our faces private, we should be much more concerned about keeping our minds private—this is, our beliefs, preferences, and goals. Using AI to pry into our minds does not require face recognition, but only the data exhaust of our mobile phones. In some of the earliest work on inferring human behavior from GPS data, my collaborators and I showed that your GPS trail reveals many of your daily activities, such as visiting friends, shopping, going to work, and so on (Liao, Fox, and Kautz 2007); see Figure 7. In 2007, smartphones with GPS were not yet on the market, so the experiments required the subjects to carry a GPS recorder. By 2012, GPS on cell phones was ubiquitous, and posts to Twitter from cell phones routinely included the user's GPS coordinates. Adam Sadilek and I showed that from location data alone one could infer with high accuracy who were both friends on Twitter and friends in real life (Sadilek and Bigham 2012). Since then, of course, user modeling from
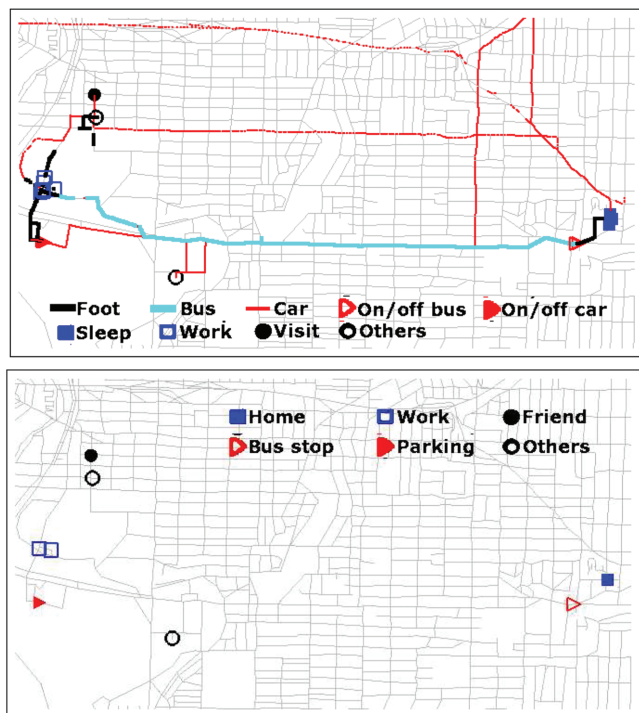
FIGURE 7 Illustration of inference on part of a GPS trace, which visited this 4 km × 2 km area several times. (a) Activities estimated for each patch. (b) Places generated by clustering significant activities, followed by a determination of place types. From Liao, Fox, and Kautz (2007)

mobile phone data has exploded, fusing data from GPS, social media posts, purchases, and in certain nations, the contents of private messages. In the United States, this technology has been metaphorically weaponized by organizations that want to win your business, your donations, or your vote. We have likely all experienced the phenomena whereby we purchase goods whose buyers tend to fall into a particular place on the political spectrum, and soon are bombarded with advertisements to give money to a political party associated with the position.

In China, the weaponization of AI for inferring one's beliefs from cell phone data is no metaphor. Millions of Uyghurs toil in concentration camps because an algorithm inferred that they were likely to hold traditional religious and social beliefs and thus be insufficiently loyal to the central government.[2] The hunt for "wrong thinkers" in totalitarian states certainly does not require the use of AI; we see it going on today in relatively low-tech fashion in North Korea, Cuba, and Russia, and historically in East Germany, the Soviet Union under Stalin, and China during the Cultural Revolution. Before the introduction of AI, however, mind-control required enormous expenditures by the state. In 1950, East Berlin's secret police, the Stasi, employed 2% of the population full-time or 6.5% of the

population if one included part-time informants (Koehler 2008).

AI, however, makes mind surveillance and social control cheap and scalable. First, far fewer humans need to be employed in data collection and analysis, as illustrated today in Xinjiang (Buckley and Mozur 2019). Second, AI allows better targeting of repressive measures. The imprisonment of a million Uyghurs, a full 10% of that ethnic group, could not be completely hidden from the world[3] and risked radicalizing ordinary citizens who were not caught up in the net. This may be, however, the last time that a nation needs to imprison such a large number of people to enforce social conformity. All and only those individuals who are truly likely to be rebellious and to recruit others to that point of view will need to be microtargeted. Further, the stick could be combined with a carrot by identifying and rewarding social influencers who side with the state. The next time Beijing decides to clamp down on one of the other fifty-odd ethnic groups in China, the repressive measures may be subtle enough to fly under the radar of the rest of the world. More generally, AI will enable totalitarian states to endure indefinitely (Minardi 2020).

AI-created fake (news | friends)

Fake news has constantly been in the news since the run-up to the 2016 presidential election. Fake news spreads through social media faster than real news (Vosoughi, Roy, and Aral 2018), and many pundits have claimed that fake news changes election outcomes.[4] When Open AI developed the statistical language model GPT-2, it announced that it would not make the model immediately public out of caution that it could be turned into a fake news generator by nefarious people (Open.ai 2019).

Fake news is actually much older than 2016—the phenomena goes back to the invention of printing and before. The so-called "blood libel" is a fake news story that originated in the 12th century and is still spread by anti-semites around the world (Soll 2016). In 1898, William Randolph Hearst's newspapers promoted the fake story that the US Battleship Maine had been sunk by Spain; Hearst's interest was in increasing newspaper sales, but the story also led to the United States starting the Spanish-American war (ibid). All of this is to say that people do not need the aid of AI to write fake news; fake news stories are effective even when they are implausible and badly written. Fake news is a social problem, but AI-powered fake news is not. There is, however, another application of language models and so-called deepfake image and video generation (Mirsky and Lee 2021) that is an existential threat to society: the creation and monetization of fake friends.

In a recent survey, 27% of American Millennials (20 and 30 somethings) said that they had no close friends, and 33% said that they were often or always lonely (Ballard 2019). What might be called a loneliness pandemic has infected the younger generation as well. Severe suicide by children aged 10–14 tripled over the last decade, and clinical depression doubled (Curtin and Heron 2019); psychologist Jean Twenge has argued that social-media obsession has damaged young people's abilities to make interpersonal connections (Twenge 2018).

The 2015 movie She is about a lonely middle-aged man who falls in love with the intelligent assistant on his cell phone. The movie is fantasy because She really is an intelligent—in fact, superintelligent—being. In real life, many people are ready to relate to the shallow AI agents that can be created today. Statistical language models are allowing such agents to become increasingly fluent, and videos of their faces and bodies generated by adversarial neural networks have nearly crossed the uncanny valley (Mori, MacDorman, and Kageki 2012). While such AIs do not have the capacity to think, feel, or understand, they will soon be able to simulate friendship—and with that, simulate compassion and love—to a degree that is good enough to satisfy damaged humans who long for companionship.

Human friendship is hard. Learning the subtle rules of effect human engagement is the major task a child faces from birth through adolescence. It requires practice, disappointment, and pain. AI friends—that is, fake friends—will be attractive to many because they will eliminate this pain.[5] We have argued that there is a growing population of lonely young people, and that they could be attracted to but damaged by AI friends. Why, however, should we fear that such fake friends would ever be made available to them? The answer is that enormous amounts of money could be made by commercializing fake friends. One of the largest and most profitable companies in the world, Facebook, says that its mission is to give people the power to build community (FaceBook 2021), but in fact is an advertising platform. Consider how much more effective advertisements would be if they were spoken to you by your best friend—not even appearing as formal advertisements, but simply as the suggestions of your bestie! FaceBook's revenue is currently $86 billion (Statistica 2021); can we doubt that the advertising revenue for fake friends would not be ten times—or a hundred times—that number? The cost, of course, is that the person who chooses the fake-friend route will never develop the skills needed to engage with other people; they may well become narcissists or sociopaths.

As with many societal trends, Japan is a bellwether of the future. The phrase "otaku" refers to adolescent through middle-aged men who having abandoned hope of finding a romantic relationship with a real woman call computer-generated characters their girlfriends (Rani 2013). While it may simply be sad when adults fixate on fake friends, it will be terrifying when they are marketed to children. In the United States, the majority of young children have their own tablets (Kabali, Irigoyen, and Nunez-Davis 2015) and infants are estimated to start handling mobile devices during the first year of life (Rideout 2017). My vision of the end of humanity is illustrated in Figure 8: the concept of an AI friend and lover from the movie She; minus superintelligence, because we are nowhere near that; plus an image from the television show "Blue's Clues" to illustrate targeting children with simulated human interaction.

Weapons of (environmental) mass destruction

If you have not seen the short video Slaugherbots produced by Future of Life Institute, you should put down this essay and go watch it immediately (FoLI 2017). The premise is that swarms of tiny intelligent drones will soon be released by unspecified nefarious parties (possibly the organizers of TED talks) to assassinate people. The video is more imaginative and entertaining than any number of full-length robot-apocalypse movies, but it is seriously meant to convince people to ban the development of autonomous weapons. In fact, the global campaign against autonomous weapons, of which Slaughterbots is but a part, did lead many nations (but not the United States or China) to call for them to be banned (Human Rights Watch 2020). In the United States, the Defense Innovation Board's AI Ethical Guidelines (Defense Innovation Board, 2019) state, "Human beings should exercise appropriate levels of judgment and remain responsible for the development, deployment, use, and outcomes of Department of Defense AI systems."

What is an "appropriate level of judgment" that human beings should always exert? To the Future of Life Institute, the line is clear: an AI algorithm should never make the decision to kill. In a war, an AI system might identify potential targets, but the decision to attack a target must be made by a human. But is this truly the most moral position to take with regard to autonomous weapons? Larry Lewis, Senior Advisor for the State Department on Civilian Protection in the Obama administration, and Member of the US Delegation for UN Deliberations on Lethal Autonomous Weapons Systems, has written, "Artificial intelligence may make weapons systems and the future of war relatively less risky for civilians than it is today" (Lewis 2020). In the "fog of war" when split-second decisions are required, human judgment is often quite poor; an AI algorithm would be less likely than a human to misidentify a wedding party as a terrorist group.

FIGURE 8 She—Bostrom superintelligence + blues clues = = fake friends starting in childhood

Apart from moral considerations, what of the existential risk of allowing the development of autonomous lethal drones? Slaughterbots take for granted that there will be no defenses against murderbots. However, just as missiles led to the development of antimissile defenses, slaughterbots will lead to the development of antislaughterbot defenses—and indeed Israel Aerospace Industries has announced the sale of dozens of its counter-UAV Drone Guard systems (Frantzman 2021). It can be further argued that drones—autonomous or not—are not particularly effective weapons of war. As I write this in July 2021, the United States is the process of withdrawing from Afganistan, having conclusively lost the war against the Taliban despite having launched 13,072 drone strikes since January 2015 (TBoIJ 2021).

The public outcry about AI-powered weapons in the war on terrorism has overshadowed the more deadly and insidious manner in which AI has supercharged nonsustainable exploitation of natural resources. Rather than resolutions against AI-powered weapons of mass destruction, it would be better if the nations of the world united against what we might call AI-driven weapons of environmental destruction. We are all too aware of how our oceans have been devastated by overfishing, garbage dumping, and global warming. Environmental damage has been hardest in the upper and middle depths of the ocean. The deepest areas are a global network of troughs that cut across the ocean floor, which formed where tectonic plates collided; the greatest of these is the Mariana trench in the Pacific, a 1500-mile long crescent located 7 miles below the surface. Aside from some filtration of plastic waste (Morelle 2019), the Mariana Trench has been unchanged for 180 million years, and is rich with yet-to-be-classified forms of life. But all this is about to change.

There is today skyrocketing demand for rare-earth minerals for batteries, solar cells, and electrics. Many of the known deposits of minerals such as lithium are either near exhaustion or are controlled by nations antagonistic to the United States.[6] Geologists have speculated that the lowest sea flows could contain enormous deposits of these minerals, and in order to reach their companies are building enormous robotic excavators that will crawl
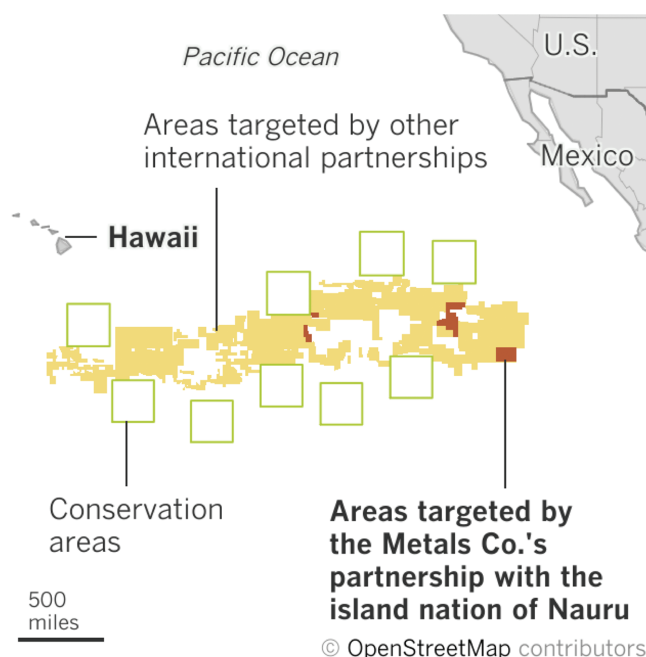

FIGURE 9 Images of mining areas for Clarion Clipperton Zone near Hawaii from Hylton (2020)

along and scape up the ocean floor. The first area where mining operations are commencing is called the Clarion Clipperton Zone off the coast of Hawaii (Figure 9). The Royal Swedish Academy of Sciences has predicted that each mining ship will release 2 million cubic feet of discharge every day (Hylton 2020). University of Hawaii oceanographer Jeff Drazen is quoted in that same article as saying, "There's a Belgian team in the CCZ doing a component test right now. They're going to drive a vehicle around on the seafloor and spew a bunch of mud up. So these things are already happening. We're about to make one of the biggest transformations that humans have ever made to the surface of the planet. We're going to strip-mine a massive habitat, and once it's gone, it isn't coming back." Another scientist, Douglas McClauley from UC Santa Barbara, is quoted in Halper (2021) as stating, "The ocean is the place on the planet where we know least about what species exist and how they function. This is like opening a Pandora's box. We're concerned this won't do much good
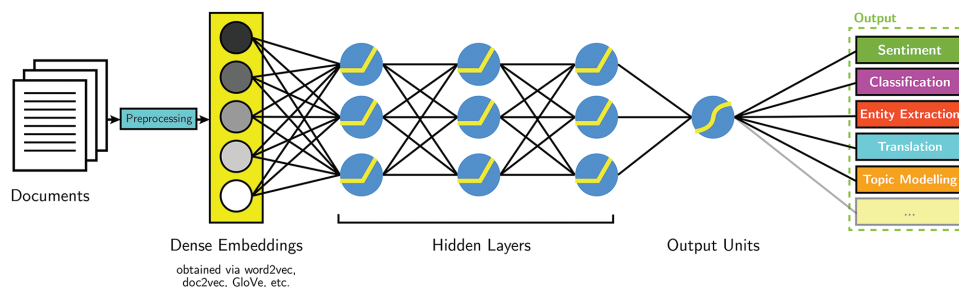
FIGURE 10   Symbolic Neuro symbolic architecture, from Aylien (2020)

for climate change, but it will do irreversible harm to the ocean."

The role of the lowest depths of the sea in earth's overall cycle of life is yet poorly understood. We do not know if the ecosystems will be resilient to mineral extraction or if their mining will trigger the collapse of the ecosystem, or what the effects such a collapse would have on life higher in the oceans and on land. It will be not a small irony if the attempt to help stop climate change by shifting to electric vehicles indirectly causes an even more deadly ecological disaster, and if the robot army that dooms the earth's biological life does not fly in the air or roll on the ground, but instead toils invisibility at the bottom of the ocean.

PART III: FUTURE OF AI

At last, we come to the crystal-ball gazing section of the essay: what will be the next big scientific advance in AI? The recent book by Gary Marcus and Ernie Davis, Rebooting AI (Marcus and Davis 2019), argues that the dominant artificial neural network approach has reached a plateau because human-level AI requires symbolic reasoning. After arguing for a return to research focused on symbolic reasoning that is reminiscent of that which flourished during the Second AI Summer, they ultimately conclude that the next advance in AI will actually be a combination of symbolic and neural net methods.

While the book makes it sound like the authors are fighting a lonely battle, they are in fact in violent agreement with deep learning researchers about the need to understand how to combine neural and symbolic approaches. A banner inscribed "Neuro-Symbolic Reasoning" could fly over all of the metaphorical armies of AI. As with so much of life, however, the devil is in details: what exactly would be the architecture of such a hybrid AI system? We will briefly survey six possible designs; for each, I have coined a name that aims to capture its essence.[7]

(1) **Symbolic Neuro symbolic** is currently the deep learning SOP (standard operating procedure) for natural language processing (Figure 10). Symbolic input—in the
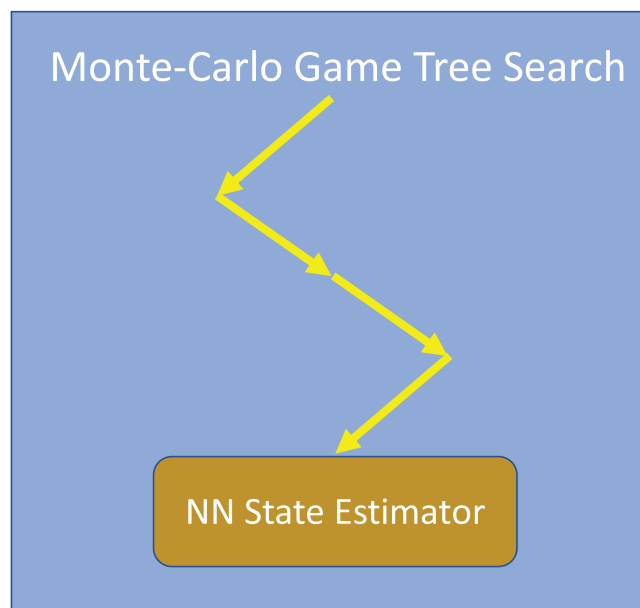

FIGURE 11   Symbolic[Neuro] architecture

case of language, sequences of words—are each converted to vectors by word2vec, GloVe, or similar (Mikolov et al. 2013; Pennington, Socher, and Manning 2014)—and passed to a neural network. The network's output units convert the previous layer to a symbolic category or sequence of symbols via a softmax operation.

(2) The **Symbolic[Neuro]** architecture employs a Neural pattern recognition subroutine within a symbolic problem solver (Figure 11). AlphaGo is a prototypical example of this design (Silver et al. 2016). The problem solver is the Monte-Carlo Tree Search algorithm Coulom (2006) and its heuristic evaluation function is a neural network. Most robots and autonomous vehicles are Symbolic[Neuro] systems.

(3) In a **Neuro | Symbolic** system, a neural network converts nonsymbolic input, such as the pixels of an image, into a symbolic data structure, which is then processed by a symbolic reasoning system (Figure 12). In the Neuro-Symbolic Concept Learner (Mao et al. 2019), the symbolic reasoning system provides a feedback signal that is used to
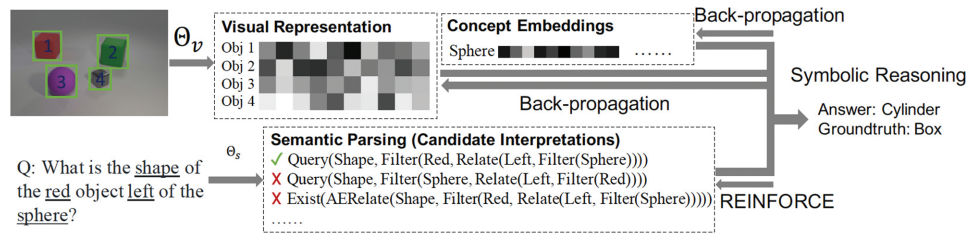
FIGURE 12 Neuro | Symbolic architecture, from Mao (Mao et al., 2019)

train the neural network. If you rotate the figure of Neuro | Symbol 90° counterclockwise, you will see that it is similar to Symbolic[Neuro]; they differ in that the neuro part is a coroutine rather than a subroutine.

(4) The **Neuro: Symbolic → Neuro** approach uses the SOP architecture but with a special training regime based on symbolic rules (Figure 13). A remarkable example is that of Lample and Charton (2020) for performing symbolic mathematics. They trained a transformer sequence to sequence deep learning system on input–output pairs of the form (A, B) where the mathematical expression A can be simplified to the expression B. After training, given a previously unseen expression, the system could usually simplify it correctly. Note that the neural network did not generate a step-by-step derivation; instead, it so thoroughly absorbed its lessons that it could simply guess the correct answer.

(5) A **Neuro_{Symbolic}** architecture transforms symbolic rules into templates for structures within the neural network (Figure 14). Tensor product representations (Smolensky et al. 2016) and logic tensor networks (Serafini, Donadello, and Garcez 2017) have been demonstrated for building abstraction and part-of hierarchies into the network. To the best of my knowledge, the approach has not been explored for encoding disjunctive rules that would enable combinatorial reasoning by cases.

(6) We finally come to the approach to neuro-symbolic reasoning that I believe has the greatest potential to combine the strengths of logic-based and neural-based AI, namely the **Neuro[Symbolic]** architecture (Figure 15). The basic idea is to embed a symbolic reasoning engine inside a neural engine, with the goal of enabling super-neuro and combinatorial reasoning. The architecture is based on Daniel Kahneman's theory of "thinking fast and slow" (2011), which states that the brain implements two distinct mechanisms for reasoning. System 1 operates automatically and quickly, with little or no effort and no sense of voluntary control, and is based on similarity. For example, you see a man frowning and conclude that he is angry, because his expression is similar to the expression on the faces of people you have seen in the past who were angry. The conclusion may be wrong—for instance, he may be frowning because of a problem with his dentures—but the

$$\text{input:} \int x^n \, dx \quad \text{output:} (1/n+1) \, x^{n+1}$$

FIGURE 13 Symbolic → Neuro architecture

system works well enough for 99% of everyday reasoning. System 2, by contrast, allocates attention to the effortful mental activities. You consciously and often painfully work your way through a tree of choices and imagined outcomes. People make errors during System 2 reasoning not because the underlying rules of thought are unsound but because the human brain is so poorly designed to do it. We overlook possible choices and miscalculate probabilities. We are almost sure to become lost if more than a dozen or so reasoning steps are required.

The properties of System 1 and System 2 are remarkably similar to those of the artificial neural net approach to AI and the logical approach to AI. Note that even when one is executing System 2, System 1 is ultimately in charge; it is System 1 that decides when to initiate System 2. The name Neuro[Symbolic] is chosen to indicate that the symbolic subsystem is a subroutine of the main neural system. A natural instantiation of the architecture is a reinforcement learning agent that includes in its set of actions one to start System 2 executing. It might also have actions that monitor time and computational resource usage by System 2 and terminate it when its resource use is excessive.

How does System 1 send a description of the problem to be solved to System 2, and how does System 2 return its answer to System 1? Since System 2 works on symbolic structures, System 1 must generate an internal symbolic representation of the task. In Rethinking Consciousness, the psychologist Graziano (2019) hypothesizes that there is a cognitive mechanism named the Attention Schema[8] with which the brain generates a symbolic representation of what is it thinking about. The System 1 action to initiate System 2 must therefore first fill the Attention Schema with a symbolic representation of the task at hand. Generating such an internal symbolic structure might use the same kind of artificial neural net structures that are used to generate sentences in natural language models. The output of the symbolic reasoning subsystem could be fed back into the neural network just as natural language is input to
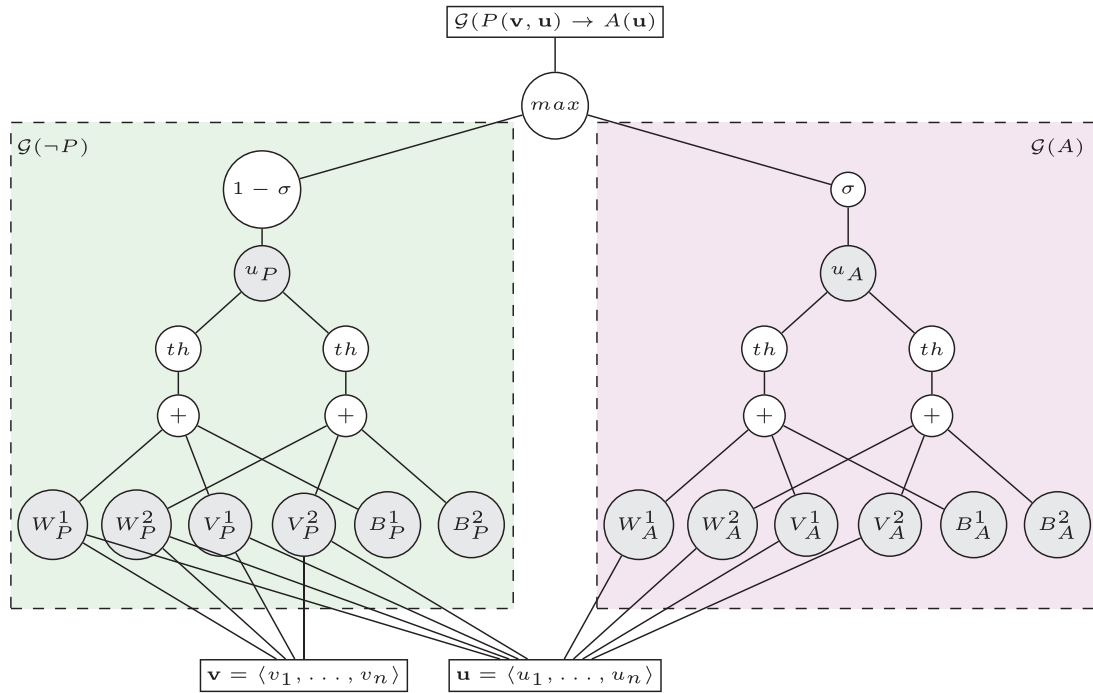
FIGURE 14 Neuro_{Symbolic} architecture, from Serafina et al. Serafini, Donadello, and Garcez 2017

a deep learning system; in other words, as a "little voice in the head." An alternative feedback mechanism would be for the output of System 2 to modify the preceptive field of System 1, as illustrated in Figure 14. The architecture can be generalized to include many different specialized symbolic reasoning subsystems, such as A* state-space search, constraint satisfaction, numeric and symbolic mathematics, and first-order theorem proving. Such subsystems can be vastly more powerful than human System 2 reasoning. For example, Heule, Kullmann, and Marek (2016) used a version of the DPLL Boolean satisfiability algorithm to solve a Pythagorean Triples Problem that required a 200 terabyte proof. No human could create such a proof.

Yoshi Bengio and collaborators (Madan et al. 2021) have proposed a related architecture called recurrent independent mechanisms that includes an Attention Schema-like module that contains a reduced vector representation of the ANN's state rather than a fully symbolic representation. We do not yet know, however, how even reduced vector representations could support fast combinatorial search. A possible objection to our proposed Neuro[Symbolic] architecture is that the symbolic solver would not in general be differentiable, so it would not support gradient-descent-based learning of the System 1 part of the system. System 1 could still be trained, however, on input/output pairs from the symbolic solver, as is done in Neuro: Symbolic → Neuro. Over time, System 1 could become better and better at predicting solutions to problem, and thus learn to invoke System 2 less frequently. This matches our intuition about how we learn subjects such as arithmetic. We begin by laboriously calculating even simple sums, but eventually learn to solve at least two-digit arithmetic problems reflectively. Another criticism of the Neuro[Symbolic] approach we outlined is that it is limited to logical reasoning or problems that can be reduced to logical reasoning. There is no fundamental reason, however, that it could not be extended to probabilistic reasoning. For example, the Attention Schema could be instantiated with Bayesian network where System 2 engine is a probabilistic reasoning engine. However, similarity reasoning, as opposed to probabilistic reasoning, would need to remain within System 1. Finally, an entirely valid criticism of Neuro[Symbolic] is that while inspired by models of cognition, it certainly does not model the brain at what Marr called the implementation level (Marr 1982) and may even differ at the algorithmic level. For example, it seems implausible that when solving a logic problem we actually perform DPLL search with clause learning, as does Heule's system (ibid). This is not a problem if our ultimate goal is not to understand human intelligence but to create AI systems that can solve the countless problems in science, engineering, and commerce that are beyond human abilities.

SUMMARY: THE STATE OF AI

The history of AI is not the cartoon version with which we began this essay. Each AI Summer has led to lasting
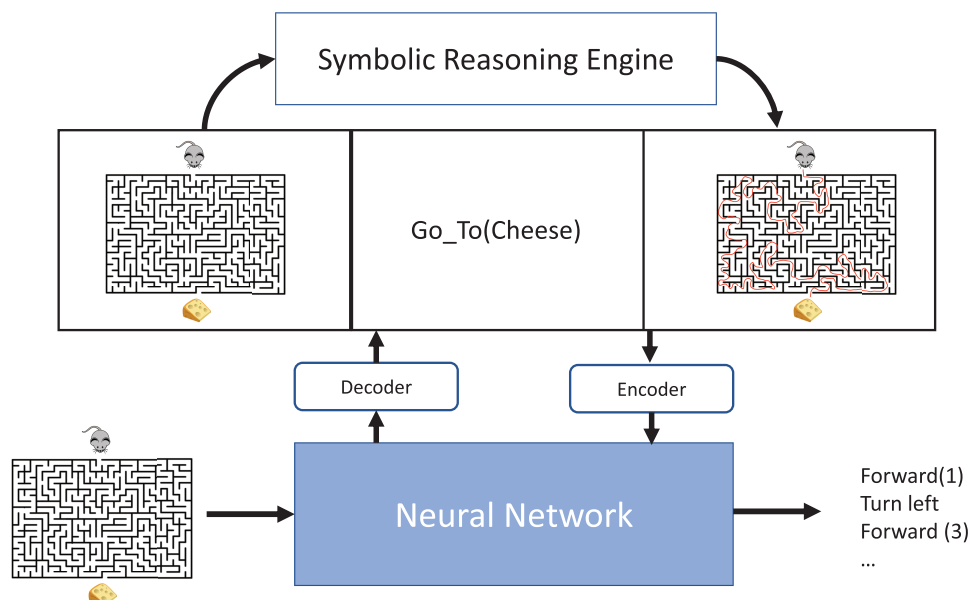
FIGURE 15 Neuro[Symbolic] architecture illustrated by a mouse-maze domain. The System 1 agent sees and recognizes a maze beyond which lies cheese, which causes it to choose the action of invoking a System 2 symbolic reasoning system—in this case, a shortest-path algorithm. A shortest path algorithm is a specific case of combinatorial search. As part of making the choice, it instantiates a grid-world version of the map in its Attention Schema. In this example, the search algorithm records its solution by making annotations to the Attention Schema: marks on the grid that show the path. The System 1 agent has learned how to interpret the marks to help guide it to the cheese

insights. Each AI Winter was caused by backlash against unfulfilled promises - unfulfilled because the methods of the time ran into technical roadblocks. Science continued on quietly during the winters, finding ways around the roadblocks and devising complete approaches and algorithms. The current AI summer is again a time when many exaggerated promises are being made, but this time the ratio of real-world results to hype is much higher than before. For all the wild claims that we are nearing General AI or the singularity—few actually made by serious researchers—as a whole the field of AI today is positively modest compared to the hype surrounding, for example, blockchain or quantum computing.[9]

AI can be used to drive powerful applications for good or for evil. Most of the world's efforts to constrain bad uses of AI are, in my opinion, misguided. They focus on improbable problems while ignoring the ways that AI is most likely to damage society, human development, and the earth itself. Remarkable real-world accomplishments have been made. For example, while I was finishing this essay, Google announced what might well be the most important practical result in the history of AI: the success of AlphaFold in predicting the 3D shape of proteins (Senior et al. 2020). Many scientists believe that the next set of scientific advances will come through the integration of neural and symbolic approaches to AI—but we do not yet know what form that integration will have.

Federal funding for fundamental research in AI has been flat for most of the past 20 years, but has recently begun to increase. In 2020, the National Science Foundation together with industry and federal agency partners funded the first cohort of National Artificial Intelligence Research Institutes, which are designed to support long-term fundamental research, promote the application of AI to problems of national importance, and grow and diversify the next generation of AI scientists and engineers. At the time of writing this essay (August 2021), Congress is considering greatly enlarging the nation's investment in AI and other technologies of the future. Although this call for increasing investment is partly driven by rivalry with China, we can at least hope that the competition will result in broad benefits to society by creating general AI technology that will be used to improve our health, our livelihoods, and our global environment. We hope that the AI race will be similar in its impacts to the space race rather than the nuclear arms race.

material are those of the author and do not reflect the views of the National Science Foundation.

CONFLICT OF INTEREST
No conflict of interest has been declared by the author(s).

ORCID
Henry A. Kautz https://orcid.org/0000-0001-5219-2970

ENDNOTES
[1] An interesting historic note is the Bram Cohen who appears as coauthor on the paper that introduced the Walksat algorithm is the same Bram Cohen who went on to invent Bittorent.

[2] At the time I first delivered the lecture on which this essay is based in January 2020, news of the camps in Xinjiang had only recently appeared in the US mainstream media, and many still claimed it was all just anti-Chinese propaganda. This was also just before COVID-19 exploded in the United States, and government officials on both sides of the Pacific Ocean were claiming that the disease was mostly contained and would fizzle out in a few weeks, so I made the trip to deliver the lecture in person in New York City. I worried that my comments in this section would turn out to be exaggerated, but had no worries about making the trip; in retrospect my concerns should have been exactly the reverse of what they were.

[3] Although when the Disney corporation filmed the remake of Mulan in Xinjiang Province, by strange change their cameras never happened to capture the concentration camps.

[4] However, a study of voters' online media consumption leading up to the 2016 election of Donald Trump by Guess, Nyhan, and Reifler (2020) found that "these (fake news) websites made up a small share of people's information diets on average and were largely consumed by a subset of Americans with strong preferences for proattitudinal information. These results suggest that the widespread speculation about the prevalence of exposure to untrustworthy websites has been overstated."

[5] This is one of the themes of the recent novel Klara and the Sun (Ishiguro 2021).

[6] For example, the largest reserve of lithium on land may be in Afganistan (Horowitz 2021).

[7] This is only one possible taxonomy of neuro-symbolic system; see Garcez and Lamb (2020) for another.

[8] Graziano's Attention Schema is not to be confused with the concept of attention in deep learning. The latter refers to various algorithms that have been proposed for combining the vector representations of sequential data points.

[9] This is not to disparage the *theory* of quantum computing, which is revealing astonishing facts about the relationship between computing and quantum physics. The problem is that much of the hype about the practical impact of quantum computing is coming from scientists who should know better (Aaronson 2021). The parallels with the hype that came from "inside the house" during the first two AI summers is striking.

REFERENCES
Aaronson, Scott. 2021. "QC Ethics and Hype: The Call is Coming From Inside the House." Shetetl-Optimized (The Blog of ScottAaronson). March 20. https://www.scottaaronson.com/blog/?p=5387. Published March 20, 2021.

Albright, Thomas D. 2017. "Why Eyewitnesses Fail." *Proceedings of the National Academy of Sciences* 114: (30): 7758–64.

Allen, James F., Philip R. Cohen, Robin Cohen, and C. Raymond Perrault, Corot Reason, and Mary. Horrigan-Tozer. 1977. "A Computer Model of Conversation." *SIGART Newsletter* (61): 29.

Aylien. 2020. "Leveraging Deep Learning for Multilingual Sentiment Analysis." last updated November 8, 2020. https://aylien.com/blog/leveraging-deep-learning-for-multilingual (accessed September 1, 2022).

Ballard, Jamie. 2019. "Millennials are the Loneliest Generation." last updated July 30, 2019. https://today.yougov.com/topics/lifestyle/articles-reports/2019/07/30/loneliness-friendship-new-friends-poll-survey

Bayardo, R. J. & R. C. Schrag. 1997. "Using CSP Look-Back Techniques to Solve Real-World SAT Instances." In Proceedings of the Fourteenth National Conference on Artificial Intelligence, Providence, RI, 203–8.

Beam, Paul, Henry Kautz, and Ashish Sabharwal. 2004. "Towards Understanding and Harnessing the Potential of Clause Learning." *Journal of Artificial Intelligence Research* 22: 319–51.

Bellman, R. 1957. "*Dynamic Programming*." Mineola, NY: Dover Publications.

Boole, George. 1854. "*The Laws of Thought*." London: MacMillan and Co.

Brachman, R. J., and James G. Schmolze. 1985. "An Overview of the KL-ONE Knowledge Representation System." *Cognitive Science* 9: 171–216.

Brigs, Rick. 1985. "Knowledge Representation in Sanskrit and Artificial Intelligence." *AI Magazine* 6 (1): 32.

Buckley, Chris, and Paul Mozur. 2019. "How China Uses High-Tech Surveillance to Subdue Minorities." *New York Times*, May 22, 2019.

Buolamwini, Joy, and Timnit Gebru. 2018. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency, PMLR*, 81, 77–91.

TBoIJ (Bureau of Investigative Journalism). Drone Statistics. Accessed August 26, 2021, https://www.thebureauinvestigates.com/projects/drone-war/.

Chinoy, Sahil. 2019. "The Racist History Behind Facial Recognition." *New York Times*.

Conger, Kate, Richard Fausset, and Serge F. Kovaleski. 2019. "San Francisco Bans Facial Recognition Technology." *The New York Times*.

Cook, Stephen. 1971. "The Complexity of Theorem Proving Procedures." In *Proceedings of the Third Annual ACM Symposium on Theory of Computing*, 151–8.

Cortes, Corinna, and V. Vapnik. 2004. "Support-Vector Networks." *Machine Learning* 20: 273–97.

Coulom, Rémi Coulom. 2006. "Efficient Selectivity and Backup Operators in Monte-Carlo Tree Search." In Proceedings of the 5th International Conference on Computers and Games, Turin, Italy. https://dl.acm.org/doi/proceedings/10.5555/1777826

Crockford, Kade. 2020. "How is Face Recognition Surveillance Technology Racist?" ACLU of Massachusetts Technology for Liberty Project, last updated June 16, 2020. https://www.aclu.org/news/privacy-technology/how-is-face-recognition-surveillance-technology-racist/

Curtin, Sally C., and Melonie Heron. 2019. "Death Rates Due to Suicide and Homicide Among Persons Aged 10–24: United States,

2000–2017." *National Center for Health Statistics Data Brief* (352): 1–8. https://www.cdc.gov/nchs/data/databriefs/db352-h.pdf

Davis, Martin, George Logemann, and Donald Loveland. 1961. "A Machine Program for Theorem Proving." *Communications of the ACM* 5 (7): 394–7.

Defense Innovation Board. 2019. "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense." last updated October 31, 2019. https://innovation.defense.gov/ai/

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data via the EM Algorithm." *Journal of the Royal Statistical Society Series B.* 39 (1): 1–38.

Devlin, J., Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the NAACL.*

Domingos, Pedro. 2015. *The Master Algorithm*. Basic Books.

Ehrlich, Pippa, and James Reed. 2020. "My Octopus Teacher. Documentary film." *Netflix.*

Facebook. 2021. "Facebook Investor Relations: FAQs." Retrieved August 20, 2021. https://investor.fb.com/resources

Feigenbaum, E. A., J. Lederber, and R. Buchanan. 1968. "Heuristic Dendral." In *Proceedings of the Hawaii International Conference on System Sciences*, University of Hawaii Press. https://en.wikipedia.org/wiki/Hawaii_International_Conference_on_System_Sciences

Ferrucci, David, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T. Mueller. 2013. "Watson: Beyond Jeopardy!." *Artificial Intelligence* 199: 93–105.

Fikes, Richard E., and Nils J. Nilsson. 1971. "Strips: A New Approach to the Application of Theorem Proving to Problem Solving." In *Proceedings of the Second International Joint Conference on Artificial Intelligence*, 1–3.

Frantzman, Seth. 2021. "Who is Buying Israeli Counter-Drone Systems in South Asia?" *Defense News.*

Friedman, N., L. Getoor, D. Koller, and A. Pfeffer. 1999. "Learning Probabilistic Relational Models." In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI).*

Fukushima, Kunihiko. 1980. "Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position." *Biological Cybernetics* 36: 193–202.

FoLI (Future of Life Institute). Slaughterbots. Youtube vido, November, 2017. https://www.youtube.com/watch?v=HipTO_7mUOw.

Garcez, Artur d'Avila, and Luis C. Lamb. 2020. "Neurosymbolic AI: The 3rd Wave." arXiv.org. last updated December 16, 2020. https://arxiv.org/abs/2012.05876

Graziano, Michael S. A. 2019. "*Rethinking Consciousness: A Scientific Theory of Subjective Experience.*" Norton.

Grother, Patrick, Mei Ngan, and Kayee Hanaoka. 2019. "Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects." NISTR 8280, National Institute of Standards and Technology. last updated 2019. http://doi.org/10.6028/NIST.IR.8280.pdf

Guess, A. M., B. Nyhan, and J. Reifler. 2020. "Exposure to Untrustworthy Websites in the 2016 US Election." *Nature Human Behavior* 4: 472–80.

Gupta, Naresh, and Dana S. Nau. 1991. "Complexity Results for Blocks-World Planning." In *Proceedings of the Ninth National Conference on Artificial Intelligence.*

Gödel, K. 1931. "Uber formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I." *Monatshefte für Mathematik Physik* 38: 173–98.

Halper, Evan. 2021. "California's Electric Car Revolution, Designed to Save the Planet, Also Unleashes a Toll on It." *Los Angeles Times.*

Hart, P. E., N. J. Nilsson, and B. Raphael. July 1968. "A Formal Basis for the Heuristic Determination of Minimum Cost Paths." *IEEE Transactions on Systems Science and Cybernetics* 4 (2): 100–7.

Hayes, Patrick J. Hayes. 1978. "The Naive Physics Manifesto." In *Expert Systems in the Micro-Electronic Age*, edited by D. Michie. Edinburgh University Press.

Heckerman, D., and Shortliffe, E. 1992. "From certainty factors to belief networks." *Artificial Intelligence in Medicine* 4: (1): 35–52.

Heule, Marijn J. H., Oliver Kullmann, and Victor W. Marek. 2016. "Solving and Verifying the Boolean Pythagorean Triples Problem Via Cube-and-Conquer." In *Theory and Applications of Satisfiability Testing: SAT 2016, Volume 9710 of Lecture Notes in Computer Science*, edited by Nadia Creignou and Daniel Le Berre, Springer, 228–45.

Hill, Kashmir Hill. 2020. "Wrongfully Accused by an Algorithm." *New York Times.*

Hoggett, Reuben. 2011. "W. Grey Walter and his Tortoises." last modified June 6, 2011. http://cyberneticzoo.com/cyberneticanimals/w-grey-walter-and-his-tortoises/ (accessed October 4, 2020).

Hylton, Wil S. 2020. "History's Largest Mining Operation is About to Begin." *The Atlantic* (January/February 2020).

Innocence Project. 2020. Eyewitness Identification Reform. retrieved August 13, 2020. https://innocenceproject.org/eyewitness-identification-reform/

Ishiguro, Kazuo. 2021. "*Klara and the Sun: A Novel.*" Knopf.

Kabali, H. K., M. M. Irigoyen, R. Nunez-Davis, et al. 2015. "Exposure and Use of Mobile Media Devices by Young Children." *Pediatrics* 136 (6): 1044–50.

Kahneman, Daniel. 2011. "*Thinking, Fast and Slow.*" Farrar, Straus and Giroux.

Kautz, Henry. 1987. "A Formal Theory of Plan Recognition." PhD Thesis, TR 215, Department of Computer Science, University of Rochester.

Kirkpatrick, S., C. D. Gelatt Jr., and M. P. Vecchi. May 1983. "Optimization by Simulated Annealing." *Science* 13: 671–80.

Koehler, John O. Koehler. 2008. "*Stasi: The Untold Story of the East German Secret Police.*" Basic Books.

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "Image Classification with Deep Convolutional Neural Networks." In *Proceedings of the Twenty-sixth Annual Conference on Neural Information Processing Systems (NIPS)*, 1106–14.

Lample, Guillaume, and François Charton. 2020. "Deep Learning For Symbolic Mathematics." In *Proceedings of the International Conference on Learning Representations.*

LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation* 1 (4): 541–51.

Lee, Honglak, Roger B. Grosse, R. Ranganath, and A. Ng. 2011. "Unsupervised Learning of Hierarchical Representations with Convolutional Deep Belief Networks." *Communications of the ACM* 54: 95–103.

Lee, Kai-Fu. 2018. "*AI Superpowers: China, Silicon Valley, And The New World Order.*" Mariner Books.

Lewis, Larry. January 10, 2020. "Killer Robots Reconsidered: Could AI Weapons Actually Cut Collateral Damage?" *Bulletin of the Atomic Scientists.*

Liao, L., D. Fox, and H. Kautz. 2007. "Extracting Places and Activities from GPS Traces Using Hierarchical Conditional Random Fields." *International Journal of Robotics Research* 26 (1): 119–34.

Lighthill, J. 1973. "Artificial Intelligence: A General Survey." In *Artificial Intelligence: A Paper Symposium.* United Kingdom: Science Research Council.

Madan, Kanika, Rosemary Nan Ke, Anirudh Goyal, B. Scholkopf, and Y. Bengio. 2021. "Fast and Slow Learning of Recurrent Independent Mechanisms." arXiv.org, last updated 2021. https://arxiv.org/abs/2105.08710

Mao, Jiayuan, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision." In *Proceedings of the International Conference on Learning Representations.*

Marcus, Gary, and Ernest Davis. 2019. "*Rebooting AI.*" Random House.

Marques-Silva, J. P., and K. A. Sakallah. November 1996. "GRASP—A New Search Algorithm for Satisfiability." In *Proceedings of the International Conference on Computer-Aid Design*, 220–7.

Marr, David. 1982. "*Vision: A Computational Investigation into The Human Representation and Processing of Visual Information.*" San Francisco: W. H. Freeman.

McCarthy, John. 1958. "Programs with Common Sense." In *Proceedings of the Symposium on Mechanization of Thought Processes*, Teddington, England, National Physical Laboratory.

McCulloch, W. S., and W. H. Pitts. 1943. "A Logical Calculus of the Ideas Immanent in Nervous Activity." *Bulletin of Mathematical Biophysics* 7: 115–33.

McDermott, John. 1980. "*R1: An Expert in the Computer Systems Domain*". Proceedings of the First AAAI Conference on Artificial Intelligence. AAAI'80. Stanford, California: AAAI Press: 269–271.

Mikolov, Tomas, Kai Chen, G. Corrado, and J. Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." In *Proceedings of the 4th International Conference on Learning Representations (ICLR 2016)*, San Juan, Puerto Rico. https://dl.acm.org/doi/proceedings/10.5555/1777826

Minardi, Di. 2020. "The Grim Fate That Could Be 'Worse Than Extinction'." *BBC*. last updated October 15, 2020. https://www.bbc.com/future/article/20201014-totalitarian-world-in-chains-artificial-intelligence

Mirsky, Yisroel, and Wenke Lee. 2021. "The Creation and Detection of Deepfakes: A Survey." *ACM Computing Surveys* 54 (1): 1–41.

Morelle, Rebecca. 2019. "Mariana Trench: Deepest-ever Sub Dive Finds Plastic Bag." *BBC News.* last updated May 13, 2019. https://www.bbc.com/news/science-environment-48230157

Mori, M., K. F. MacDorman, and Norri Kageki. 2012. "The Uncanny Valley." *IEEE Robotics and Automation* 19 (2): 98–100.

Muggleton, S., and C. Feng. 1990. "Efficient Induction of Logic Programs." In *Proceedings of the 19th International Conference on Algorithmic Learning Theory (ALT)*, 368–81.

Newell, A., J. C. Shaw, and H. A. Simon. 1959. "Report on a General Problem-Solving Program." In *Proceedings of the International Conference on Information Processing*, 256–64.

Newell, Alan, and Herbert Simon. 1956. "The logic theory machine: A complex information processing system." *IRE Transactions on Information Theory* 2: 61–79.

Newell, Allan, and Herbert A. Simon. 1972. "*Human Problem Solving.*"

Open.ai. 2019. "*Better Language Models and Their Implications.*" last updated February 14, 2019. https://openai.com/blog/better-language-models

Pearl, Judea. 1988. "*Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.*" Morgan Kaufmann.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar. https://dl.acm.org/doi/proceedings/10.5555/1777826

Pople, Harry E. Jr. 1976. "Presentation of the INTERNIST System." In *Proceedings of the Artificial Intelligence in Medicine Workshop*, New Brunswick, N.J: Rutgers University.

Pushak, Yasha & Holger Hoos. 2020. "Advanced Statistical Analysis of Empirical Performance Scaling." In Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), July 8–12, 2020. New York, NY: Association for Computing Machinery.

Quinlan, J. R. 1986. "Induction of Decision Trees." *Machine Learning* 1 (1): 81–106.

Radford, Alec, Jeffrey Wu, Rewon Cild, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." last updated 2018. https://openai.com/blog/better-language-models/

Rani, Anita. 2013. "The Japanese Men Who Prefer Virtual Girlfriends to Sex." *BBC News*, last updated October 24, 2013. https://www.bbc.com/news/magazine-24614830

Richardson, Matthew, and Pedro Domingos. 2006. "Markov Logic Netwoks." *Machine Learning* 62 (1–2): 107–36.

Richens, Richard H. 1956. "Preprogramming for Mechanical Translation." *Mechanical Translation* 3 (1): 20–5.

Rideout, V. 2017. "*The Common Sense Census: Media Use by Kids Age Zero to Eight.*" San Francisco, CA: Common Sense Media.

Rosenblatt, F. 1958. "The perceptron: A probabilistic Model for Information Storage and Organization in the Brain." *Psychological Review* 65 (6): 386–408.

Rumelhart, D., G. Hinton, and R. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 323: 533–6.

Russakovsky, Olga, J. Deng, Hao Su, J. Krause, S. Satheesh, S. Ma, Zhiheng Huang, A. Karpathy, A. Khosla, Michael S. Bernstein, A. Berg, and Fei-Fei Li. 2015. "ImageNet Large Scale Visual Recognition Challenge." *International Journal of Computer Vision* 115: 211–52.

Russell, B. 1903. "*The Principles of Mathematics.*"

Russell, Stuart. 2019. "*Human Compatible.*" Penguin.

Sadilek, Adam Henry Kautz, and Jeffrey P. Bigham. 2012. "Finding Your Friends and Following Them to Where You Are." In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining (WSDM)*, Seattle, WA. https://dl.acm.org/doi/proceedings/10.5555/1777826

Selman, Bart, Hector Levesque, and David Mitchell. 1992. "A New Method for Solving Hard Satisfiability Problems." In *Proceedings of the Tenth National Conference on Artificial Intelligence.*

Selman, Bart, Henry Kautz, and Bram Cohen. 1996. "Local Search Strategies for Satisfiability Testing." DIMACS Series in Discrete Mathematics and Theoretical Computer Science 26.

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek,

Alexander W R Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, and David Silver. 2020. "Koray Kavukcuoglu, and Demis Hassabis. "Improved Protein Structure Prediction using Potentials from Deep Learning." *Nature* 577 (7792): 706–10.

Serafini, Luciano, Ivan Donadello, and Artur Garcez. 2017. "Learning and Reasoning in Logic Tensor Networks: Theory and Application to Semantic Image Interpretation." In *Proceedings of the Symposium on Applied Computing (SAC)*, 125–30, Marrakech, Morocco: ACM Special Interest Group on Applied Computing (SIGAPP). April 3–7.

Shortliffe, Edward H., and Bruce G. Buchanan. 1975. "A Model of Inexact Reasoning in Medicine." *Mathematical Biosciences* 23 (3–4): 351–79.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Chrittwieser, Ionnis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timonthy Lillicrap, Madeleine Leach, and Koray Kavukcuoglu. 2016. "Thore Graepel, Thore and Demis Hassabis. "Mastering the Game of Go with Deep Neural Networks and Tree Search." *Nature* 529 (7587): 484–9.

Singhal, Amit. 2012. "Introducing the Knowledge Graph: Things, Not Strings." last updated May 16, 2012, retrieved September 6, 2014. https://blog.google/products/search/introducing-knowledge-graph-things-not/

Smolensky, P., Moontae Lee, X. He, Wen-tau Yih, Jianfeng Gao, and L. Deng. 2016. "Basic Reasoning with Tensor Product Representations." arXiv.org last updated 2016. https://arxiv.org/abs/1601.02745

Soll, Jacob. 2016. "The Long and Brutal History of Fake News." *Politico*. last updated December 18, 2016. https://www.politico.com/magazine/story/2016/12/fake-news-history-long-violent-214535

Statistica. 2021. "Facebook's Annual Revenue from 2009 to 2020." last updated February 5, 2021. https://www.statista.com/statistics/268604/annual-revenue-of-facebook/

Sutton, R. S., and A. G. Barto. 1981. "Toward a Modern Theory of Adaptive Networks: Expectation and Prediction." *Psychological Review* 88 (2): 135–70.

Sutton, Richard S. 1988. "Learning to Predict by the Methods of Temporal Differences." *Machine Learning* 44: 3–9.

Twenge, Jean M. 2018. "*Why Today's Super-Connected Kids Are Growing Up Less Rebellious, More Tolerant, Less Happy—and Completely Unprepared for Adulthood—and What That Means for the Rest of Us.*" Atria Books.

Valiant, L. 1984. "A Theory of the Learnable." In *Proceedings of the Sixteenth Annual ACM Symposium on Theory of Computing (STOC)*.

Vitti, Joseph J. 2013. "Cephalopod Cognition in an Evolutionary Context: Implications for Ethology." *Biosemiotics* 6: 393–401.

Vosoughi, Soroush, Deb Roy, and Sinan Aral. 2018. "The Spread of True and False News Online." *Science* 359: 1146–51.

Walter, W. Grey, R. Cooper, V. J. Aldridge, W. C. McCallum, and A. L. Winter. 1964. "Contingent negative variation: An electrical sign of sensorimotor association and expectancy in the human brain." *Nature* 203: 380–4.

Walter, William Grey. 1953. "*The Living Brain.*" Duckworth.

Weiner, Norbert. 1948. "*Cybernetics: Or Control and Communication in the Animal and the Machine.*" MIT Press.

Werbos, Paul. 1974. "Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences." PhD thesis, New York: Harvard University Reprinted in the book:, and Paul J. Werbos 1994. The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting.: John Wiley & Sons.

Whitehead, Alfred North, and Bertrand Russell. 1910–1913. "*Principia Mathematica.*"

Witten, Ian H. August 1977. "An Adaptive Optimal Controller for Discrete-Time Markov Environments." *Information and Control* 34 (4): 286–95.

Zadeh, L. A. 1965. "Fuzzy sets." *Information and Control* 8 (3): 338–53.

AUTHOR BIOGRAPHY

Henry Kautz is a Professor in the Department of Computer Science and was the founding director of the Goergen Institute for Data Science at the University of Rochester. He served as a Division Director for Information & Intelligent Systems (IIS) at the National Science Foundation from 2018 to 2021, where he led the National AI Research Institutes program. He has been a researcher at AT&T Bell Labs in Murray Hill, NJ, and a full professor at the University of Washington, Seattle. In 2010, he was elected President of the Association for Advancement of Artificial Intelligence (AAAI), and in 2016 was elected Chair of the American Association for the Advancement of Science (AAAS) Section on Information, Computing, and Communication. His interdisciplinary research includes practical algorithms for solving worst-case intractable problems in logical and probabilistic reasoning; models for inferring human behavior from sensor data; pervasive healthcare applications of AI; and social media analytics. In 1989 he received the IJCAI Computers & Thought Award, which recognizes outstanding young scientists in artificial intelligence, and 30 years later received the 2018 ACM-AAAI Allen Newell Award for career contributions that have breadth within computer science and that bridge computer science and other disciplines. At the 2020 AAAI Conference, he received both the Distinguished Service Award and the Robert S. Engelmore Memorial Lecture Award.