A Large List of Neural Network Literature

This completely arbitrary and unscientific list of literature related to the general domain of Artificial Neural Networks lists interesting and sometimes obscure works that picked my fancy. I have added notable events and works from the more general domain of “AI” (in particular, symbolic AI) to have a way of putting Artificial Neural Networks in their historical context. The more we get to the present, the more technical the literature becomes, as would be expected

📆	Event
🚀	Work of Fiction
📔	Book
📘	Overview article
🐿️	Popular Science Article

1800s		
1834+ 📆	Charles Babbage tries to design a general-purpose computing engine: the "Analytical Engine".	After being unable to finish the "Difference Engine" due to perfectionism, Charles Babbage conceives the idea of the "Analytical Engine". While the "Difference Engine" was designed as a special-purpose calculator (interpolate polynomials via the method of finite differences and print the result), the "Analytical Engine" would have been a generic "stored program computer", with the program stored on punched card as used in the Jacquard Loom. It was never built.
1847	**Méthode générale pour la résolution des systèmes d’équations simultanées** Augustin-Louis Cauchy C. R. Acad. Sci. Paris, 25:536–538, 1847. https://cs.uwaterloo.ca/~y328yu/classics/cauchy-en.pdf	🦉 **Introduction of Gradient Descent**
1863	**Darwin among the Machines** Samuel Butler https://en.wikipedia.org/wiki/Darwin_among_the_Machines https://archive.org/details/cu31924013448299/page/n67/mode/2up?q=darwin	🦉 **The base of Frank Herbert's word construction "Butlerian Jihad".** Wikipedia states: *"Darwin among the Machines" is a letter to the editor published in The Press newspaper on 13 June 1863 in Christchurch, New Zealand. The title, which was chosen by the author, references the work of Charles Darwin. Written by Samuel Butler but signed Cellarius, the letter raised the possibility that machines were a kind of "mechanical life" undergoing constant evolution, and that eventually machines might supplant humans as the dominant species (...) The letter ends by urging that, "War to the death should be instantly proclaimed against them. Every machine of every sort should be destroyed by the well-wisher of his species. Let there be no exceptions made, no quarter shown; let us at once go back to the primeval condition of the race."*
1887+ 📆	Santiago Ramón y Cajal becomes professor in Barcelona and begins studying neural tissue under the microscope, staining individual neurons and tracing their connections. Cajal establishes the **neuron doctrine**, which states that the brain is made up of discrete nerve cells (neurons) that communicate via synapses.	🦉 **Serious neuro-physiology begins** https://en.wikipedia.org/wiki/Santiago_Ram%C3%B3n_y_Cajal
1930s		
1933	**Grundbegriffe der Wahrscheinlichkeitsrechnung** Andrei Kolmogoroff Berlin, Verlag von Julius Springer 1933 https://archive.org/details/kolmogoroff-1933-grundbegriffe-der-wahrscheinlichkeitsrechnung	🦉 **Andrei Kolmogorov creates a mathematically sound axiomatic basis for any calculus of probabilities** *See also:* **The origins and legacy of Kolmogorov’s Grundbegriffe** Glenn Shafer, Vladimir Vovk https://arxiv.org/abs/1802.06071 April 25, 2003, marked the 100th anniversary of the birth of Andrei Nikolaevich Kolmogorov, the twentieth century's foremost contributor to the mathematical and philosophical foundations of probability. The year 2003 was also the 70th anniversary of the publication of Kolmogorov's *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Kolmogorov's *Grundbegriffe* put probability's modern mathematical formalism in place. It also provided a philosophy of probability - an explanation of how the formalism can be connected to the world of experience. In this article, we examine the sources of these two aspects of the Grundbegriffe - the work of the earlier scholars whose ideas Kolmogorov synthesized.
1936	**On Computable Numbers, with an Application to the Entscheidungsproblem**	🦉 **"Computation" set on a firm theoretical base**

	https://londmathsoc.onlinelibrary.wiley.com/doi/abs/10.1112/plms/s2-42.1.230 (paywalled!)	Alan Turing defines what should be understood by a "effective computation procedure" (as proposed by Hilbert) by conceiving of the **Turing Machine**: a finite state machine working with a potentially infinite external memory in the form of a tape. This both a highly intuitive understanding of what could be meant by "effective computation procedure" and also provides the theoretical base of of understanding what computers can actually do. It can be easily shown that there is a set of special machines, the **Universal Turing Machine**, able to emulate all the other Turing Machines, hence the concept of the **general purpose computer**. Earlier proposals for defining "effective computation procedure", Gödel-Kleene-Herbrand **general recursive functions** and Alonzo Church's **Lambda Calculus**, are quickly shown to be equivalent to the Turing Machine concept. All have the same "power of computation", i.e. can "effectively compute" the same set of functions.
1940s		
1943	**A logical calculus of the idea immanent in nervous activity** Warren S. McCulloch, Walter Pitts (University of Illinois, College of Medicine) Appears in: Bulletin of Mathematical Biophysics, Vol. 5, 1943 Reprinted in: Bulletin of Mathematical Biophysics, Vol. 52, 1990 https://link.springer.com/article/10.1007/BF02478259 (paywalled)	See also: **The Man Who Tried to Redeem the World with Logic** "Walter Pitts rose from the streets to MIT, but couldn't escape himself." http://nautil.us/issue/21/information/the-man-who-tried-to-redeem-the-world-with-logic by Amanda Gefter
1946-03 🚀	**A Logic Named Joe** Will F. Jenkins alias Murray Leinster Appears in: "Astounding Science Fiction" n°184 https://www.isfdb.org/cgi-bin/title.cgi?865228 https://en.wikipedia.org/wiki/A_Logic_Named_Joe	A "terminal" computer starts to ransack databases and delivers information about anything to anyone who queries it over the telecommunication network, endangering society. The "logic repairman" saves the day by locating it and replacing it by another terminal from storage.
1948	**Intelligent Machinery** Alan M. Turing Report, National Physics Laboratory, Mathematics Division https://ia801703.us.archive.org/23/items/turing1948/turing1948_text.pdf	The possible ways in which machinery might be made to show intelligent behaviour are discussed. The analogy with the human brain is used as a guiding principle. It is pointed out that the potentialities of the human intelligence can only be realized if suitable education is provided. The investigation mainly centres round an analogous teaching process applied to machines. The idea of an unorganized machine is defined, and it is suggested that the infant human cortex is of this nature. Simple examples of such machines are given, and their education by means of rewards and punishments is discussed. In one case the education process is carried through until the organization is similar to that of an ACE.
1948	**A Mathematical Theory of Communication** Claude Shannon Appears in: "The Bell System Technical Journal" Vol. 27, July, October, 1948. https://archive.org/details/ost-engineering-shannon1948/mode/2up	🦉 **Claude Shannon introduces Information Theory** See also: "Shannon's Theorem - Math 280 Notes" https://www.collegesidekick.com/study-docs/6049584 For a note on the notion of "entropy", see: https://www.quantamagazine.org/what-is-entropy-a-measure-of-just-how-little-we-really-know-20241213/ and also: **Researchers in an Entropy Wonderland: A Review of the Entropy Concept** Marko Popovic https://arxiv.org/abs/1711.07326
1948 📔	**Cybernetics: Or Control and Communication in the Animal and the Machine** Norbert Wiener https://en.wikipedia.org/wiki/ Cybernetics:_Or_Control_and_Communication_in_the_Animal_and_the_Machine	🦉 **The classical postwar text that defined the new branch of "Cybernetics"** For its genesis, based on Norbert Wiener having the leeway of visiting off-limits research labs during WWII, see David A. Mindell: *"Between Human and Machine: Feedback, Control, and Computing before Cybernetics"*
1950s		
1950	**Computing Machinery and Intelligence** Alan M. Turing Appears in: "Mind", Volume LIX, Issue 236, October 1950, Pages 433–460 https://courses.cs.umbc.edu/471/papers/turing.pdf	🦉 **Introduces the idea of "The Imitation Game" (aka. "The Turing Test")** An idea open to much criticism. Wikipedia lists the latter here: https://en.wikipedia.org/wiki/Computing_Machinery_and_Intelligence
1951-09	**A stochastic approximation method** Herbert Robbins, Sutton Monro. Appears in: Annals of Mathematical Statistics 22, 3 (1951), 400–407. https://doi.org/10.1214/aoms/1177729586	🦉 **Introduces stochastic gradient descent**
1954 📆	"Theory of Neural-Analog Reinforcement Systems and Its Application to the Brain Model Problem"	Marvin Minsky completes his PhD Thesis at Princeton University under the supervision of John McCarthy and Albert W. Tucker.
1956 📆	An **eight weeks summer workshop takes place at Dartmouth College**, Hanover, New Hampshire from about June 18 to August 17. The project was formally proposed to the Rockefeller Foundation in 1955 to obtain financing by McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon. The proposal is credited	20 people are listed as having attended at some point in Solomonoff's notes (according to Wikipedia), but apparently this list is still incomplete: • Ray Solomonoff, American mathematician • Marvin Minsky, American cognitive and computer scientist

	with introducing the term 'artificial intelligence'. https://en.wikipedia.org/wiki/Dartmouth_workshop https://home.dartmouth.edu/about/artificial-intelligence-ai-coined-dartmouth https://spectrum.ieee.org/dartmouth-ai-workshop (May 2023, Grace Solomonoff) https://spectrum.ieee.org/history-of-ai (September 2021, Eliza Strickland)	• John McCarthy, American cognitive and computer scientist • Claude Shannon, American mathematician, electrical engineer, computer scientist • Trenchard More, American mathematician and computer scientist • Nathaniel Rochester, Chief architect of the IBM 701, • Oliver G. Selfridge, English mathematician and computer scientist • Julian Bigelow, American computer engineer and cybernetician • William Ross Ashby, English psychiatrist and a pioneer in cybernetics • Warren Sturgis McCulloch, American neuropsychologist and cybernetician • Abraham Robinson, German mathematician • Tom Etter, American computer scientist • John Forbes Nash, American mathematician • David Sayre, American scientist • Arthur Lee Samuel, American computer scientist • Kenneth Radford Shoulders, American experimental physicist • "Shoulders' friend" • Alex Bernstein, American mathematician and computer scientist • Herbert Simon, American scholar on computer science, economics, and cognitive psychology. • Allen Newell, American researcher in computer science and cognitive psychology Additionally, an article in IEEE Spectrum by Grace Solomonoff identifies Peter Milner on a photo.
1957-01	**The Perceptron: A Perceiving and Recognizing Automaton (Project PARA)** Frank Rosenblatt Cornell Aeronautical Laboratory, Inc. Report No. 85–460–1.	🦉 **Introduction of the "Perceptron", a machine learning algorithm performing two-class linear regression on a space of features.** Available via this page: https://websites.umass.edu/brain-wars/1957-the-birth-of-cognitive-science/the-perceptron-a-perceiving-and-recognizing-automaton
1958	**The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain** Frank Rosenblatt, Cornell Aeronautical Laboratory Appears in: "Psychological Review", Vol. 65, No. 6, 19S8 http://deeplearning.cs.cmu.edu/S24/document/readings/Rosenblatt_1959-09865-001.pdf	
1958	**Pandemonium: a paradigm for learning** Oliver G. Selfridge Appears in: "Mechanisation of Thought Processes: Proceedings of a Symposium Held at the National Physical Laboratory, November 1958, London", HMSO, pp. 513-526 https://gwern.net/doc/ai/nn/1959-selfridge.pdf	🦉 **Introduction of the "Pandemonium" concept** We are proposing here a model of a process which we claim can adaptively improve itself to handle certain pattern recognition problems which cannot be adequately specified in advance. Such problems are usual when trying to build a machine to Imitate any one of a very large class of human data processing techniques. A speech typewriter is a good example of something that very many people have been trying unsuccessfully to build for some time. We do not suggest that we have proposed a model which can learn to typewrite from merely hearing speech. Pandemonium does not, however, seem on paper to have the same kinds of inherent restrictions or inflexibility that many previous proposals have had. The basic motif behind our model is the notion of parallel processing. This is suggested on two grounds: first, it is often easier to handle data in a parallel manner, and, indeed, it is usually the more "natural" manner to handle it in; and, secondly, it is easier to modify an assembly of quasi-independent modules than a machine all of whose parts interact immediately and in a complex way. We are not going to apologize for a frequent use of anthropomorphic or biomorphic terminology. They seem to be useful words to describe our notions. What we are describing is a process, or, rather, a model of a process. We shall not describe all the reasons that led to Its particular formulation, but we shall give some reasons for hoping that it does in fact possess the flexibility and adaptability that we ascribe to it.
1958-11 📆	**"Symposium on the Mechanization of Thought Processes" in Teddington, Surrey** A 4-day conference conference, entiltled "Mechanization of Thought Processes", is held at the National Physics Laboratory in Teddington, Surrey. Over 200 people attend. Two volumes of proceedings are published in 1959. A review by "Nature", January 24. 1959: https://www.nature.com/articles/183225a0.pdf "Mechanization of Thought Processes", Volume 1 and 2, can be obtained at the Internet Archive: https://archive.org/details/mechanisationoft01nati https://archive.org/details/mechanisationoft02nati	See: **The Representation of Knowledge and the Relevance of Biological Models at the Symposium on the Mechanization of Thought Processes, 1958** by Matthew Cobb appears in: "IEEE Annals of the History of Computing", July/September 2023 https://ieeexplore.ieee.org/document/10190123 “Mechanization of Thought Process” was an international conference involving researchers from academia, government, industry, and the military that took place in the U.K. in 1958. It saw the first presentation of McCarthy’s *Advice Taker* and of Selfridge’s *Pandemonium*, and one of the first expositions of Rosenblatt’s Perceptron, as well as presentations on new programming languages, cybernetic experiments, and simple diagnostic systems. This article describes the conference and the occasionally boisterous debates that took place, drawing out the common challenges faced by researchers at the time, focusing on the relevance of biological models for mechanized systems of thought processing and the difficulty of embodying knowledge or context in a system to enable it to solve problems effectively. Particular attention is paid to the methodological criticisms of work in both machine translation and in what we would now consider to be artificial intelligence made by the Israeli linguist and philosopher Yehoshua Bar-Hillel.

1960s		
1960 📆	The first "International Congress on Automatic Control" sponsored by IFAC (International Federation of Automatic Control) is held in Moscow from June 27 to July 2, hosted by Alexander Letov, President of IFAC at that time. A rare Cold War-era event where Eastern and Western scientists came together to exchange ideas on automation and cybernetics.	For some anecdotes, see: **Recollections of Norbert Wiener and the First IFAC World Congress** by Bernard Widrow (as told to Barbara Field) https://isl.stanford.edu/~widrow/papers/j2001recollectionsof.pdf from "IEEE Control Systems Magazine", June 2001
1961 📙	**Computers and Common Sense: The Myth of Thinking Machines** Mortimer Taube	🦉 **Mortimer Taube criticizes the current "AI" programme. See:** https://ieeexplore.ieee.org/document/9324949 **The "General Problem Solver" Does Not Exist: Mortimer Taube and the Art of AI Criticism** Shunryu Colin Garvey, Stanford University Institute for Human-Centered AI, USA This article reconfigures the history of artificial intelligence (AI) and its accompanying tradition of criticism by excavating the work of Mortimer Taube, a pioneer in information and library sciences, whose magnum opus, *Computers and Common Sense: The Myth of Thinking Machines* (1961), has been mostly forgotten. To convey the essence of his distinctive critique, the article focuses on Taube's attack on the general problem solver (GPS), the second major AI program. After examining his analysis of the social construction of this and other "thinking machines," it concludes that, despite technical changes in AI, much of Taube's criticism remains relevant today. Moreover, his status as an "information processing" insider who criticized AI on behalf of the public good challenges the boundaries and focus of most critiques of AI from the past half-century. In sum, Taube's work offers an alternative model from which contemporary AI workers and critics can learn much.
1965	**Cybernetic Predicting Devices.** Alexey Ivakhnenko, Valentin G. Lapa CCM Information Corporation.	🦉 **Earliest "deep learning" algorithm** From Jürgen Schmidhuber's "Annotated History of Modern AI and Deep Learning": Successful learning in deep feedforward network architectures started in 1965 in the Ukraine (back then the USSR) when Alexey Ivakhnenko and Valentin Lapa introduced the first general, working learning algorithms for deep MLPs (multi-layer perceptrons) with arbitrarily many hidden layers (already containing the now popular multiplicative gates). A paper of 1971 (Ivakhnenko, A. G. (1971). *Polynomial theory of complex systems*. IEEE Transactions on Systems, Man and Cybernetics, (4):364-378.) already described a deep learning net with 8 layers, trained by their highly cited method which was still popular in the new millennium, especially in Eastern Europe, where much of Machine Learning was born. See: https://ieeexplore.ieee.org/document/4308320
1966 🚀	**Destination: Void** Frank Herbert https://www.isfdb.org/cgi-bin/title.cgi?2253 https://en.wikipedia.org/wiki/Destination%3A_Void	A classic "AI breakout" novel by Frank Herbert. Notable as it ruminates on a way to imbue the AI ethical principles before it reaches superhuman (and supernatural) capabilities. Still hard to read.
1968-07 📙	**Semantic Information Processing** Marvin Minsky, Editor MIT Press	A classic collection of texts from the bronze age of symbolic AI.
1968 🚀	**2001 - A Space Odyssey** Novelization of Stanley Kubrick's movie by Arthur C. Clarke For an elaboration of Clarke and Kubrick's collaborative work on this project, see *The Lost Worlds of 2001*, Arthur C. Clarke, Signet, 1972. https://www.isfdb.org/cgi-bin/title.cgi?2485 https://en.wikipedia.org/wiki/2001:_A_Space_Odyssey_(novel) Note that Marvin L. Minsky was hired by Kubrick to consult on HAL 9000.	**Chapter 16 - Hal** Hal (for Heuristically programmed ALgorithmic computer, no less) was a masterwork of the third computer breakthrough. These seemed to occur at intervals of twenty years, and the thought that another one was now imminent already worried a great many people. The first had been in the 1940s, when the long-obsolete vacuum tube had made possible such clumsy, high-speed morons as ENIAC and its successors. Then, in the 1960s, solid-state microelectronics had been perfected. With its advent, it was clear that artificial intelligences at least as powerful as Man's need be no larger than office desks – if one only knew how to construct them. Probably no one would ever know this; it did not matter. In the 1980s, Minsky and Good had shown how neural networks could be generated automatically – self replicated – in accordance with any arbitrary learning program. Artificial brains could be grown by a process strikingly analogous to the development of a human brain. In any given case, the precise details would never be known, and even if they were, they would be millions of times too complex for human understanding. Whatever way it worked, the final result was a machine intelligence that could reproduce – some philosophers still preferred to use the word "mimic" – most of the activities of the human brain – and with far greater speed and reliability. It was extremely expensive, and only a few units of the HAL9000 series had yet been built; but the old jest that it would always be easier to make organic brains by unskilled labor was beginning to sound a little hollow.
1969 📙	**Perceptrons: An Introduction to Computational Geometry** Marvin L. Minsky, Seymour A. Papert (dedicated to psychologist Frank Rosenblatt) MIT Press	https://en.wikipedia.org/wiki/Perceptrons_(book) • Second printing with handwritten notes, 1972 • Expanded edition, 1988 • Expanded edition reprint 2017, foreword by Léon Bottou Wikipedia states: These perceptrons were modified forms of the perceptrons introduced by Rosenblatt in 1958. They consisted of a retina, a single layer of input functions and a single output.

		Besides this, the authors restricted the "order", or maximum number of incoming connections, of their perceptrons. Sociologist Mikel Olazaran explains that Minsky and Papert "maintained that the interest of neural computing came from the fact that it was a parallel combination of local information", which, in order to be effective, had to be a simple computation. To the authors, this implied that "each association unit could receive connections only from a small part of the input area"· Minsky and Papert called this concept "conjunctive localness"
1970		
1970	**The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors.** Seppo Linnainmaa, Master's Thesis (in Finnish), Univ. Helsinki, 1970. https://people.idsia.ch/~juergen/linnainmaa1970thesis.pdf *See chapters 6-7 and FORTRAN code on pages 58-60.*	🦉 **"Backpropagation" algorithm first published** From Jürgen Schmidhuber's "Annotated History of Modern AI and Deep Learning": *In 1970, Seppo Linnainmaa was the first to publish what's now known as backpropagation, the famous algorithm for credit assignment in networks of differentiable nodes,[BP1,4,5] also known as "reverse mode of automatic differentiation." It is now the foundation of widely used NN software packages such as PyTorch and Google's Tensorflow.*
1971-10	**Polynomial Theory of Complex Systems** A. G. Ivakhnenko, Cybernetic Center of NASU, Kiev, Ukraine Appears in: "IEEE Transactions on Systems, Man, and Cybernetics" (Volume: SMC-1, Issue: 4, October 1971) https://ieeexplore.ieee.org/document/4308320	🦉 **Describes a deep learning net with 8 layers** A complex multidimensional decision hypersurface can be approximated by a set of polynomials in the input signals (properties) which contain information about the hypersurface of interest. The hypersurface is usually described by a number of experimental (vector) points and simple functions of their coordinates. The approach taken in this paper to approximating the decision hypersurface, and hence the input-output relationship of a complex system, is to fit a high-degree multinomial to the input properties using a multilayered perceptronlike network structure. Thresholds are employed at each layer in the network to identify those polynomials which best fit into the desired hypersurface. Only the best combinations of the input properties are allowed to pass to succeeding layers, where more complex combinations are formed. Each element in each layer in the network implements a nonlinear function of two inputs. The coefficients of each element are determined by a regression technique which enables each element to approximate the true outputs with minimum mean-square error. The experimental data base is divided into a training and testing set. The training set is used to obtain the element coefficients, and the testing set is used to determine the utility of a given element in the network and to control overfitting of the experimental data. This latter feature is termed "decision regularization.
1975 📆	**"Künstliche Intelligenz Rundbriefe"** In West Germany, the AI specialist group in the German Informatics Society starts to circulate the "KI-Rundbrief", an influential AI newsletter that runs until 1987.	See: **The “KI-Rundbrief,” Its Editors, and Its Community: A Perspective on West German AI, 1975–1987** Dinah Pfau; Helen Piel; Florian Müller; Jakob Tschandl; Rudolf Seising The “KI-Rundbrief” (AI newsletter) of the AI specialist group in the German Informatics Society is considered central to the emerging West German AI community. It was mailed out between 1975 and 1987, when it was turned into the journal Künstliche Intelligenz. Despite its presumed centrality, it has not been studied in detail. This article combines a quantitative analysis of which research was published in the newsletter heuristically with a qualitative analysis looking at the role of the newsletter and its editors. It focuses especially on the subject areas deduction, natural language processing, expert systems, image processing, and cognitive science, showing that the newsletter only partially represents artificial intelligence research in the Federal Republic of Germany. Appears in: IEEE Annals of the History of Computing (Volume: 45, Issue: 3, 01 July-Sept. 2023) https://ieeexplore.ieee.org/document/10192322
1976 🚀	**The Computer as a Symbol of God: Ellison's Macabre Exodus** Charles J. Brady Appears in: "JGE: The Journal of General Education", Vol. 28, No. 1, Spring, 1976, pp. 49–62. http://www.jstor.org/stable/27796553	**Introduction:** Computers and religion have been sharing the bed lately in some interesting variations. Authors have had computers cross check and compare the dogmas and rituals of the various world faiths to come up with a pragmatic religious formula that would appeal to the majority of mankind [Terry Carr: "Changing of the Gods", 1973], interpret the mysterious castings of the I Ching [Gregg Williams: "The Computer and the Oriental", 1973] and provide the answer to a maiden's prayer [Leonard Tushnet: "Matchmaker, Matchmaker", 1971]. A Vatican computer was found favoring the election of a robot pope [Robert Silverberg: "Good News from the Vatican", 1971], and an ultra sophisticated computer became so miffed at being asked whether or not God exists it deliberately gave a wrong answer [Barry N. Malzberg: "A Short Religious Novel"]. Two short stories in particular, "The Monster in the Clearing" [1971] by Michael Fayette and "I Have No Mouth and I Must Scream" [1971] by Harlan Ellison, deserve close analysis. Both use the computer as a symbol of God and make sharp religious statements reminiscent of the intensity and depth of the "death-of-God" theologians.
1980s		
1980-04	**Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position.** Kunihiko Fukushima appears in: "Biological Cybernetics" Volume 36, pages 193–202 https://link.springer.com/article/10.1007/BF00344251 (paywalled) Presentation on YouTube: https://youtu.be/Bh5uPyerI1M "02.Kunihiko Fukushima: Artificial Vision by Deep CNN Neocognitron"	🦉 **Convolutional NN for optical character recognition** A neural network model for a mechanism of visual pattern recognition is proposed in this paper. The network is self-organized by “learning without a teacher”, and acquires an ability to recognize stimulus patterns based on the geometrical similarity (Gestalt) of their shapes without affected by their positions. This network is given a nickname “neocognitron”. After completion of self-organization, the network has a structure similar to the hierarchy model of the visual nervous system proposed by Hubel and Wiesel. The network consits of an input layer (photoreceptor array) followed by a cascade connection of a number of modular structures, each of which is composed of two layers of cells connected in a cascade. The first layer of each module consists of “S-cells”, which show characteristics similar to simple cells or lower order hypercomplex cells, and the second layer consists of “C-cells” similar to complex cells or higher order hypercomplex cells. The afferent synapses to each S-cell have plasticity and are modifiable. The network has an ability of unsupervised learning: We do not need any “teacher” during the process of self-organization, and it is only needed to present a set of stimulus patterns repeatedly to the input layer of the network. The network has been simulated on a digital computer. After repetitive presentation of a set of stimulus patterns, each stimulus pattern has become to elicit an output only from one of the C-cell of the last layer, and conversely, this C-cell has become selectively responsive only to that stimulus pattern. That is, none of the C-cells of the last layer responds to more than one stimulus pattern. The response of the C-cells of the last layer is not affected by the pattern's position at all. Neither is it affected by a small change in shape nor in size of the stimulus pattern.
1981-10 📆	Japan's Ministry of International Trade and Industry (MITI) announces the "5th	Prolog (at that point still in infancy) is the language chosen as the pivotal element

	Generation Computer Project", soon causing a bit of a panic in the USA about a soon-to-happen "AI Gap" and demands to invest in equivalent projects ASAP.	See: https://en.wikipedia.org/wiki/Fifth_Generation_Computer_Systems
1983 📆	DARPA starts the counter to MITI's "5th Generation Computer Project", the "Strategic Computing Initiative"	LISP is the language chosen as pivotal element as the US AI community doesn't particularly like Prolog See: https://en.wikipedia.org/wiki/Strategic_Computing_Initiative Wikipedia writes: The goal of SCI, and other contemporary projects, was nothing less than full machine intelligence. "The machine envisioned by SC", according to Alex Roland and Philip Shiman, "would run ten billion instructions per second to see, hear, speak, and think like a human. The degree of integration required would rival that achieved by the human brain, the most complex instrument known to man." ... By the late 1980s, it was clear that the project would fall short of realizing the hoped-for levels of machine intelligence. Program insiders pointed to issues with integration, organization, and communication. When Jack Schwarz ascended to the leadership of IPTO in 1987, he cut funding to artificial intelligence research (the software component) "deeply and brutally", "eviscerating" the program (wrote Pamela McCorduck). Schwarz felt that DARPA should focus its funding only on those technologies which showed the most promise. In his words, DARPA should "surf", rather than "dog paddle", and he felt strongly AI was not "the next wave". ... The project was superseded in the 1990s by the Accelerated Strategic Computing Initiative and then by the Advanced Simulation and Computing Program. These later programs did not include artificial general intelligence as a goal, but instead focused on supercomputing for large scale simulation, such as atomic bomb simulations.
1983-09	**Introduction to the fifth generation (cover article)** Pamela McCorduck appears in: "Communications of the ACM" Vol. 26, No. 9 (September 1983) https://cacm.acm.org/issue/september-1983/ https://dl.acm.org/doi/10.1145/358172.358177	*Pamela McCorduck writes:* In October 1981, Japan's Ministry of International Trade and Industry (MITI) sponsored a conference to announce a new national project. Alongside national projects in supercomputing and robotics, there would be an effort to develop a new generation (the fifth, by their reckoning) of computers. he Fifth Generation will not be traditional computers. Instead, they'll be symbolic inference machines, capable of reasoning their way swiftly through massive amounts of knowledge and data. They'll be computers that can learn, associate, make inferences, make decisions, and otherwise behave in ways usually considered the exclusive province of human reason. Even their name signals the change: knowledge information processing systems, or KIPS. KIPS will be the engine of the information society; small, robust and inexpensive. They will appear as universal appliances, as commonplace and easy to use as the telephone. The project's ten-year plan is divided into three successive stages. The first three-year stage is devoted to the development of a prototype machine, a personal PROLOG workstation that will have a knowledge base comparable to present-day expert systems (thousands of rules and thousands of objects) but whose reasoning powers will be a million logical inferences per second (LIPS), an order of magnitude improvement over software-based PROLOG implementations on today's common mainframe computers such as the DEC 2060. The prototype should be finished sometime in 1984, with commercial products due a year or so later. This first phase is Japan's opportunity to climb the learning curve, and is explicitly planned for that purpose. The second four-year stage is for engineering experimentation, prototyping, continuing experiments at significant applications, and the initial experiments at systems integration. The first thrust at the major problems of parallel processing will be done in those years. The final three-year phase will concentrate on advanced engineering, building the final major engineering prototypes, and further systems integration work. The ultimate goal, scheduled for the early 1990s, is nothing less than an inference supercomputer, capable of a million to a billion LIPS, with a knowledge base that can handle tens of thousands of inference rules and hundreds of millions of objects--about the right size to encompass the Encyclopedia Britannica. The Japanese will rely heavily on bootstrapping: the project's earlier work on CAD will be used in later hardware design, for example.
1984-09	**Computer Recreations: Perceptrons** A.K. Dewdney Appears in: "Scientific American", Column "Computer Recreations", 1984-09 https://www.scientificamerican.com/issue/sa/1984/09-01/	Explains the how the Perceptron (in this case a neural network with two processing layers, with neurons of the first layer being arbitrary maps of "small regions of pixels" to the labels 0 or 1, which seems rather powerful and the second layer being a threshold neuron) cannot properly decide whether a figure is connected or not. Also introduces the idea of the "cat detection" device. *Note that this is 4 years after the presentation of the Neocognitron.*
1984-11 📆	**Japan's 5th Generation Computing Project closes out its "1st Phase".**	*William M. Raike writes in BYTE Magazin in April 1985:* The International Conference on Fifth Generation Computer Systems was held in Tokyo in early November 1984. Over 900 participants, representing 32 countries, each paid about $400 for the privilege of attending. Researchers from all over the world presented 62 research papers, and there were a couple of interesting panel discussions in addition to a highly informative nontechnical lecture by Ezra Vogel ("The Changing Nature of Information Societies"). The objectives of the conference were twofold: to provide an international forum for the exchange of information, ideas, and research in areas related to "new-generation computing:' and to present to the world a summary of the achievements of the Japanese government-sponsored Institute for New Generation Computer Technology (ICOT) at the end of the first of its three phases. The second phase of the 10-year project will last four years and the third phase, three years. *and ends with:* Next month I'll tell you about IBM Japan's test production of I-megabit dynamic RAM chips and about several new computers on the market here.
1986 🐿️	**The Society of Mind** Marvin Minsky Simon & Schuster	At https://en.wikipedia.org/wiki/Society_of_Mind we read: The work, which first appeared in 1986, was the first comprehensive description of Minsky's "society of mind" theory, which he began developing in the early 1970s. It is composed of 270 self-contained essays which are divided into 30 general chapters. The book was not written to prove anything specific about AI or cognitive science, and does not reference physical brain structures. Instead, it is a collection of ideas about how the mind and thinking work on the conceptual level.

1986-07	**Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Psychological and Biological Models (2 Volumes)** James L. McClelland, David E. Rumelhart, PDP Research Group The MIT Press • https://direct.mit.edu/books/monograph/4424/Parallel-Distributed-Processing-Volume • https://direct.mit.edu/books/monograph/5670/Parallel-Distributed-Processing-Volume	What makes people smarter than computers? These volumes by a pioneering neurocomputing group suggest that the answer lies in the massively parallel architecture of the human mind. They describe a new theory of cognition called connectionism that is challenging the idea of symbolic computation that has traditionally been at the center of debate in theoretical discussions about the mind. The authors' theory assumes the mind is composed of a great number of elementary units connected in a neural network. Mental processes are interactions between these units which excite and inhibit each other in parallel rather than sequential operations. In this context, knowledge can no longer be thought of as stored in localized structures; instead, it consists of the connections between pairs of units that are distributed throughout the network. Volume 1 lays the foundations of this exciting theory of parallel distributed processing, while Volume 2 applies it to a number of specific issues in cognitive science and neuroscience, with chapters describing models of aspects of perception, memory, language, and thought.
1986-07	**Learning representations by back-propagating errors** David E. Rumelhart, Geoffrey E. Hinton, Ronald J. Williams Appears in: "Nature" 323, October 1986	🦉 **Backpropagation is independently rediscovered for the domain of neural networks** We describe a new learning procedure, back-propagation, for networks of neurone-like (sic) units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure.
1987	**Kolmogorov's Mapping Neural Network Existence Theorem** Robert Hecht-Nielsen, Hecht-Nielsen Neurocomputer Corporation Appears in: "Proceedings of the IEEE First International Conference on Neural Networks (San Diego, CA)" https://cs.uwaterloo.ca/~y328yu/classics/Hecht-Nielsen.pdf	🦉 **You just need 2 layers with weak conditions on the (unknown) activation functions to implement any continuous function from the n-dimensional cube to m-dimensional real space exactly with 2*n+1 units in layer 1 and m units in layer 2** An improved version of Kolmogorov's powerful 1957 theorem concerning the representation of arbitrary continuous functions from the n-dimensional cube to the real numbers in terms of one dimensional continuous functions is reinterpreted to yield an existence theorem for mapping neural networks. Dedicated to Andrei Nikolaevic Kolmogorov. *Note this is the basis of the physics-oriented Kolmogorov-Arnold Networks.* *The paper concludes:* **Implications for Neurocomputing** Kolmogorov's Mapping Neural Network Existence Theorem is a statement that our quest for approximations of functions by networks is, at least in theory, sound. The above form of the Kolmogorov theorem is particularly nice since it clearly identifies all of the ingredients (as opposed to other forms of the theorem in which the existence of the required constants and/or functional expressions is established but their exact form is not specified). The direct usefulness of this result is doubtful, at least in the near term, because no constructive method for developing the g_i (transfer functions of the output layer) is known . However, it is likely that more will be learned about this form ·of mapping network in the years to come. A potentially high-payoff challenge is to discover an adaptive mechanism whereby the g_i 's could self-organize themselves in response to incoming example x and y vector pairs .
1987-01	**Neural networks, Part 1: What are they and why is everybody so interested in them now?** Appears in: "IEEE Expert" (Volume: 2, Issue: 4, January 1987) https://ieeexplore.ieee.org/document/5006524 (paywalled) "IEEE Expert" is a predecessor magazine of "IEEE Intelligent Systems"	An extremely short introduction to multilayer networks, but only the feed-forward computation is discussed, no learning or back-propagation is discussed here. This issue also features "**Neural Networks- Conference update and overview**" by Lance B. Eliot where we read: *Why did the neural network cross the road? To make a few connections on the other side. The road led to San Diego last June, however, where the IEEE First Annual International Conference on Neural Networks took place. Approximately 2000 attendees learned about the latest neural "wetware" (this is of course not the right term, ed.) during the four-day inaugural. Some readers may be unfamiliar with neural networks (a current inside joke is - Have you heard about the top-selling AWAT neural net package? Every time someone says "a neural networks package," someone else says, "a what?"). For such readers, let's begin with background information. Following that summary, I'll comment on the San Diego conference and speculate on the emerging field of neural networks.* *(...)* *Numerous computer vendors promoted their hard/soft/wet/neurowares at the conference. One notable firm, the Hecht-Nielsen Neurocomputer Corporation of San Diego, offers an elaborate range of products. HNC displayed one such product - ANZA (covered in this issue's Products section), a coprocessor system for the IBM AT and compatibles that combines specialized hardware and software for neural network processing. Among other firms included were Nestor Inc. (Rhode Island), Verac Inc. (San Diego), AIWARE Inc. (Cleveland), Neural Systems Inc. (Vancouver), NCI (New Jersey), Neuraltech Inc. (Portola Valley), Neuronics Inc. (Chicago), SAIC (Tucson), plus power hitters such as Texas Instruments and TRW.*
1987-09	**Intelligence without representation** Rodney A. Brooks Appears in: "Artificial Intelligence" 47 (1991), 139–159 https://www.sciencedirect.com/science/article/abs/pii/000437029190053M https://people.csail.mit.edu/brooks/papers/representation.pdf	Artificial intelligence research has foundered on the issue of representation. When intelligence is approached in an incremental manner, with strict reliance on interfacing to the real world through perception and action, reliance on representation disappears. In this paper we outline our approach to incrementally building complete intelligent Creatures. The fundamental decomposition of the intelligent system is not into independent information processing units which must interface with each other via representations. Instead, the intelligent system is decomposed into independent and parallel activity producers which all interface directly to the world through perception and action, rather than interface to each other particularly much. The notions of central and peripheral systems evaporate -everything is both central and peripheral. Based on these principles we have built a very successful series of mobile robots which operate without supervision as Creatures in standard office environments.
1988 📔	**Perceptrons: An Introduction to Computational Geometry** **Expanded Edition** (first edition 1969) Marvin L. Minsky, Seymour A. Papert MIT Press	A review of the "expanded edition" by Stephen Grossberg can be found in "AI Magazine" Volume 10, Number 2 (1989): https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/download/748/666 where we read: *The expanded edition of Perceptrons (MIT Press, Cambridge, Mass, 1988, 292 pp , $12.50) by Marvin L. Minsky and Seymour A. Papert comes at a time of unprecedented interest in the biological and technological modeling of neural networks. One need only to refer to the August 1988 issue of AI Expert and Philip Chapnick's editorial noting that more than 50 percent of the Fortune 500 companies are investigating neural network technologies.*

1988-04	**Undebuggability and Cognitive Science** Christopher Cherniak Appears in: "Communications of the ACM", April 1988 https://cacm.acm.org/research/undebuggability-and-cognitive-science/	A resource-realistic perspective suggests some indispensable features for a computer program that approximates all human mentality. The mind's program would differ fundamentally more from familiar types of software. These features seem to exclude reasonably establishing that a program correctly and completely models the mind.
1988-05	**Neural networks. II. What are they and why is everybody so interested in them now?** P.D. Wasserman, T. Schwartz Appears in: "IEEE Expert" (Volume: 3, Issue: 1, Spring 1988) https://ieeexplore.ieee.org/document/2091 (paywalled) "IEEE Expert" is a predecessor magazine of "IEEE Intelligent Systems"	The learning ability of neural networks and their ability to generalize and to abstract or generate ideals from an imperfect training set are examined. Their potential for multiprocessing is considered. A brief history of neural network research is followed by a discussion of their architectures and a presentation of several specific architectures and learning techniques. The Cauchy machine, which represents a possible solution to the local minima problem encountered with virtually every other neural network training algorithm, is described. The outlook for neural nets is briefly considered. The article ends with: *Future outlook: Neural networks is a rediscovered field experiencing an explosive growth in research and application interest. Algorithms and architectures proliferate. Claims and counterclaims fill the literature. And the press produces stories at a rapid pace. Despite its longevity, neural network theory and technology is rudimentary. We find many more questions than answers, and technical knowledge remains narrowly disseminated. Numerous demonstration programs exist, but not one proven commercial application for neural networks (although software and hardware producers vaguely allude to several). The situation resembles the laser's when it was introduced. The laser had such unique properties that many people felt it must be of immense value Nevertheless, to develop even a small percentage of its commercial potential required nearly a decade. If this analogy is valid, some time will pass before neural networks find applications where their unique characteristics make them the clear method of choice. Meanwhile, all parties - researchers. commercial firms, and the press - must understand the risks of promising more than can be delivered.*
1988-07	**There exists a neural network that does not make avoidable mistakes** Ronals Gallant, Halbert White https://ieeexplore.ieee.org/document/23903 Appears in: "IEEE 1988 International Conference on Neural Networks"	🦉 **On the computational power of 2-layer NNs** The authors show that a multiple-input, single-output, single-hidden-layer feedforward network with (known) hardwired connections from input to hidden layer, monotone squashing at the hidden layer and no squashing at the output embeds as a special case a so-called Fourier network, which yields a Fourier series approximation properties of Fourier series representations. In particular, approximation to any desired accuracy of any square integrable function can be achieved by such a network, using sufficiently many hidden units. In this sense, such networks do not make avoidable mistakes.
1988-09 📙	**Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference** Judea Pearl Morgan Kaufmann Publishers Inc. Full text available at: https://dl.acm.org/doi/book/10.5555/534975	From the Publisher: *Probabilistic Reasoning in Intelligent Systems* is a complete and accessible account of the theoretical foundations and computational methods that underlie plausible reasoning under uncertainty. The author provides a coherent explication of probability as a language for reasoning with partial belief and offers a unifying perspective on other AI approaches to uncertainty, such as the Dempster-Shafer formalism, truth maintenance systems, and nonmonotonic logic. The author distinguishes syntactic and semantic approaches to uncertainty and offers techniques, based on belief networks, that provide a mechanism for making semantics-based systems operational. Specifically, network-propagation techniques serve as a mechanism for combining the theoretical coherence of probability theory with modern demands of reasoning-systems technology: modular declarative inputs, conceptually meaningful inferences, and parallel distributed computation. Application areas include diagnosis, forecasting, image interpretation, multi-sensor fusion, decision support systems, plan recognition, planning, speech recognitionin short, almost every task requiring that conclusions be drawn from uncertain clues and incomplete information. *Probabilistic Reasoning in Intelligent Systems* will be of special interest to scholars and researchers in AI, decision theory, statistics, logic, philosophy, cognitive psychology, and the management sciences. Professionals in the areas of knowledge-based systems, operations research, engineering, and statistics will find theoretical and computational tools of immediate practical use. The book can also be used as an excellent text for graduate-level courses in AI, operations research, or applied probability.
1989 📙	**Building Large Knowledge-Based Systems** Representation and Inference in the Cyc Project Douglas B. Lenat, R. V. Guha Addison-Wesley Publishing Company, Inc.	🦉 **A book explaining the Cyc "commonsense reasoning" project, a purely "symbolic reasoning" system** From a review by John F. Sowa, February 1992: http://www.jfsowa.com/pubs/CycRev93.pdf *The Cyc project, started by Doug Lenat at MCC in 1984, is the most ambitious knowledge representation project ever undertaken. It embodies Lenat's current ideas for a system intended to encode all of commonsense knowledge. By the year 1999, he hopes that "no one would even think of buying a computer that doesn't have Cyc running on it". The book by Lenat and Guha is a report on the project as it was in 1989. A review of that book must distinguish four different things: the book itself, the Cyc project as it was when the book was written, the Cyc project today, and the developments that the designers are planning for the future. Of these four, the last two are probably the most interesting.*
1989-05	**Multilayer Feedforward Networks are Universal Approximators** Kurt Hornik, Maxwell Stinchcombe, Halbert White https://www.sciencedirect.com/science/article/abs/pii/0893608089900208 See also this page which lists a lot more: https://en.wikipedia.org/wiki/Universal_approximation_theorem	🦉 **On the computational power of Deep NNs** This paper rigorously establishes thut standard multilayer feedforward networks with as few as one hidden layer using arbitrary squashing functions are capable of approximating any Borel measurable function [~ continuous] from one finite dimensional space to another to any desired degree of accuracy, provided sufficiently many hidden units are available. In this sense, multilayer feedforward networks are a class of universal approximators.
1990s		
1990	**Logical vs. Analogical or Symbolic vs. Connectionist or Neat vs. Scruffy** Marvin Minsky Appears in: "Artificial Intelligence at MIT., Expanding Frontiers", Patrick H. Winston (Ed.), Vol 1, MIT Press, 1990. Reprinted in AI Magazine, 1991 https://www.mit.edu/~dxh/marvin/web.media.mit.edu/~minsky/papers/	Engineering and scientific education conditions us to expect everything, including intelligence, to have a simple, compact explanation. Accordingly, when people new to AI ask "What's AI all about," they seem to expect an answer that defines AI in terms of a few basic mathematical laws. Today, some researchers who seek a simple, compact explanation hope that systems modeled on neural nets or some other connectionist idea will quickly overtake more traditional systems based on symbol manipulation. Others believe that symbol manipulation, with a history that goes back millennia, remains the only viable approach.

	SymbolicVs.Connectionist.html	Minsky subscribes to neither of these extremist views. Instead, he argues that Artificial Intelligence must employ many approaches. Artificial Intelligence is not like circuit theory and electromagnetism. AI has nothing so wonderfully unifying like Kirchhoff's laws are to circuit theory or Maxwell's equations are to electromagnetism. Instead of looking for a "Right Way," Minsky believes that the time has come to build systems out of diverse components, some connectionist and some symbolic, each with its own diverse justification. Minsky, whose seminal contributions in Artificial Intelligence are established worldwide, is one of the 1990 recipients of the prestigious Japan Prize---a prize recognizing original and outstanding achievements in science and technology.
1990	**Elephants Don't Play Chess** Rodney A. Brooks Appears in: "Robotics and Autonomous Systems" 6 (1990) 3-15 https://people.csail.mit.edu/brooks/papers/elephants.pdf	🦉 **Seminal paper of the "Nouvelle AI" / "New AI" rejecting symbolic processing approaches** There is an alternative route to Artificial Intelligence that diverges from the directions pursued under that banner for the last thirty some years. The traditional approach has emphasized the abstract manipulation of symbols, whose grounding, in physical reality has, rarely been achieved. We explore a research methodology which emphasizes ongoing physical interaction with the environment as the primary source of constraint on the design of intelligent systems. We show how this methodology has recently had significant successes on a par with the most successful classical efforts. We outline plausible future work along these lines which can lead to vastly more ambitious systems. Keywords: "Subsumption architecture", "Nouvelle AI"
1991 📙	**Introduction to the Theory of Neural Computation** John Hertz, Anders Krogh, Richard G. Palmer "A Lecture Notes Volume in the Santa Fe Institute Studies in the Sciences of Complexity" Santa Fe Institute #1 Addison-Wesley, 1991 https://archive.org/details/introductiontoth00hert	The book I used for self-study when taking my first course on NN. Contains: • The Hopfield Model • Extension of the Hopfield Model • Optimization Problems • Simple Perceptrons • Multi-Layer Networks • Recurrent Networks • Unsupervised Hebbian Learning • Unsupervised Competitive Learning • Formal Statistical Mechanics of Neural Networks • Appendix: Statistical Mechanics

1993 📙	**Machine Learning - An Artificial Intelligence Approach**
	Editors: R.S. Michalski, J.G. Carbonell, T.M. Mitchell
	with Contributions by J.Anderson, R.Banerji, G.Bradhaw, J.Carbonell, T.Dietterich, N.Hass, F.Hayes-Roth, G.Hendrix, P.Langley, D.Lenat, R.Michalski, T.Mitchell, J.Mostow, B.Nudel, M.Rychener, R.Quinlan, H.Simon, D.Lseeman, R.Stepp, P.Utgoff
		• General Issues in Machine Learning
		    • An Overview of Machine Learning (Jaime HG. Carbonell, Ryszard S. Michalski, and Tom M. Mitchell)
		    • Why Should Machines Learn? (Herbert A. Simon)
		• Learning From Examples
		    • A Comparative Review of Selected Methods for Learning from Examples (Thomas G. Dietterich and Ryszard S. Michalski)
		    • A Theory and Methodology of Inductive Learning (Ryszard S. Michalski)
		• Learning in Problem-Solving and Planning
		    • Learning by Analogy: Formulation and Generaléizing Plans from Past Experience (Jaime G. Carbonell)
		    • Learning by Experimentation: Acquiring and Refining Problem-Solving Heuristics (Tom M. Mitchell, Paul E. Utgoff, and Ranan Banerji)
		    • Acquisition of Proof Skills in Geometry (John R. Anderson)
		    • Using Proofs and Refutations to Learn from Experience (Frederick Hayes-Roth)
		• Learning from Observation and Discovery
		    • The Role of Heuristics in Learning by Discovery: Three Case Studies (Dounglas B. Lenat)
		    • Rediscovering Chemistry With the BACON System (Pat Langley, Gary L. Bradshaw, and Herbert A. Simon)
		    • Learning from Observation: Conceptual Clustering (Ryszard S. Michalski and Robert E. Stepp)
		• Learning from Instruction
		    • Machine Transformation of Advice into a Heuristic Search Procedure (David Jack Mostow)
		    • Learning by Being Told: Acquiring Knowledge for Information Management (Norm Haas and Gary G. Hendrix)
		    • The Instructible Production System: A Retrospective Analysis (Michael D. Rychener)
		• Applied Learning System
		    • Learning efficient Classification Procedures and their Application to Chess End Games (J. Ross Quinlan)
		    • Inferring Student Models for Intelligent Computer-Aided Instruction (Derek H. Sleeman)

1993-03 📆	The March 1993 edition of "Communications of the ACM" (Vol. 36, Num. 3) contains several articles doing a review of now terminated Japan's "5th Generation" Project
	"In summary, it can be said that ICOT has built a bridge between parallel computers and AI applications. However, with the two ends of the bridge being (perhaps temporarily) out of favor, and the bridge itself being weaker than it might be, it is perhaps too soon to expect the inauguration of the bridge to be greeted with
		• The Fifth Generation Project: Personal Perspectives (Ehud Shapior, David H.D. Warren)
		• Launching the New Era
		    • Kazuhiro Fuchi, ICOT Research Center
		    • Robert Kowalski, Imperial College
		    • Koichi Furukawa, Kelo University
		    • Kazunori Ueda, NEC Corporation

	great acclaim."	• Ken Kahn, Xerox PARC • Takashi Chikayama, ICOT Research Center • Evan Tick, University of Oregon • Epilogue (Edhud Shapiro, David H.D: Warren)
1993-06	**Multilayer feedforward networks with a nonpolynomial activation function can approximate any function** Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, Shimon Schocken https://www.sciencedirect.com/science/article/abs/pii/S0893608005801315	🦉 **On the computational power of Deep NNs, again** Several researchers characterized the activation function under which multilayer feedforward networks can act as universal approximators. We show that most of all the characterizations that were reported thus far in the literature are special cases of the following general result: A standard multilayer feedforward network with a *locally bounded piecewise continuous* activation function can approximate any continuous function to any degree of accuracy if and only if the network's activation function is not a polynomial. We also emphasize the important role of the threshold, asserting that without it the last theorem does not hold. *See also:* 1989-05: "Multilayer Feedforward Networks are Universal Approximators" (Hornik, Stinchcombe, White)
1995 📙	**The Handbook of Brain Theory and Neural Networks** Edited by Michael A. Arbib The MIT Press	A real humdinger of a book! A second edition was edited in 2002
1995 📙	**Empirical Methods for Artificial Intelligence** Book by MIT Press Paul R. Cohen https://mitpress.mit.edu/9780262534178/empirical-methods-for-artificial-intelligence/	An introduction to statistics in the context of evaluating whether an "AI" algorithm works and how well it does so.
1995-03	**Temporal Difference Learning and TD-Gammon** Gerald Tesauro Appears in: "Communications of the ACM, Volume 38, Issue 3" (March 1995) https://dl.acm.org/doi/10.1145/203330.203343	🦉 First successful application of Reinforcement Learning via Self-Play to master Backgammon No neural networks are being used here.
1995-11	**CYC: A Large-Scale Investment in Knowledge Infrastructure** Douglas B. Lenat https://dl.acm.org/doi/10.1145/219717.219745	By codifying reams of commonsense knowledge, CYC automates the white space in documents to help standardize—and make more efficient—information retrieval, integration, and consistency checking.
1997-11	**Long Short-Term Memory** Sepp Hochreiter, Jürgen Schmidhuber Appears in: Neural Computation (Volume: 9, Issue: 8, 15 November 1997) https://ieeexplore.ieee.org/abstract/document/6795963	🦉 **Solved the vanishing gradient problem in recurrent neural networks (RNNs) by introducing memory cells with gates.** 🦉 **Introduced the forget gate (added later by Gers et al., 1999), allowing LSTMs to regulate memory updates dynamically.** 🦉**Enabled long-term dependencies to be learned effectively, unlike traditional RNNs.** Learning to store information over extended time intervals by recurrent backpropagation takes a very long time, mostly because of insufficient, decaying error backflow. We briefly review Hochreiter's (1991) analysis of this problem, then address it by introducing a novel, efficient, gradient based method called long short-term memory (LSTM). Truncating the gradient where this does not do harm, LSTM can learn to bridge minimal time lags in excess of 1000 discrete-time steps by enforcing constant error flow through constant error carousels within special units. Multiplicative gate units learn to open and close access to the constant error flow. LSTM is local in space and time; its computational complexity per time step and weight is O. 1. Our experiments with artificial data involve local, distributed, real-valued, and noisy pattern representations. In comparisons with real-time recurrent learning, back propagation through time, recurrent cascade correlation, Elman nets, and neural sequence chunking, LSTM leads to many more successful runs, and learns much faster. LSTM also solves complex, artificial long-time-lag tasks that have never been solved by previous recurrent network algorithms.
1997-07 🚀	**Jipi and the Paranoid Chip** Neal Stephenson Appears in: "Forbes magazine" (July 7, 1997) https://en.wikipedia.org/wiki/Jipi_and_the_Paranoid_Chip Full text: https://web.archive.org/web/20060830131222/http://www.vanemden.com/books/neals/jipi.html	A sci-fi short story by Neal Stephenson presenting the interesting idea of a LLM-level "paranoid" chatbot, reachable by text communications, connected to a car bomb. And it must be convinced to not not fire it.
1998 📙	**Talking Nets: An Oral History of Neural Networks** MIT Press edited by James A. Amderson and Eward Rosenfeld https://direct.mit.edu/books/book/4886/Talking-NetsAn-Oral-History-of-Neural-Networks	**Surprising tales from the scientists who first learned how to use computers to understand the workings of the human brain.** Since World War II, a group of scientists has been attempting to understand the human nervous system and to build computer systems that emulate the brain's abilities. Many of the early workers in this field of neural networks came from cybernetics; others came from neuroscience, physics, electrical engineering, mathematics, psychology, even economics. In this collection of interviews, those who helped to shape the field share their childhood memories, their influences, how they became interested in neural networks, and what they see as its future. The subjects tell stories that have been told, referred to, whispered about, and imagined throughout the history of the field. Together, the interviews form a *Rashomon*-like web of reality. Some of the mythic people responsible for the foundations of modern brain theory and cybernetics, such as Norbert Wiener, Warren McCulloch,

		and Frank Rosenblatt, appear prominently in the recollections. The interviewees agree about some things and disagree about more. Together, they tell the story of how science is actually done, including the false starts, and the Darwinian struggle for jobs, resources, and reputation. Although some of the interviews contain technical material, there is no actual mathematics in the book. **Contributors** James A. Anderson, Michael Arbib, Gail Carpenter, Leon Cooper, Jack Cowan, Walter Freeman, Stephen Grossberg, Robert Hecht-Neilsen, Geoffrey Hinton, Teuvo Kohonen, Bart Kosko, Jerome Lettvin, Carver Mead, David Rumelhart, Terry Sejnowski, Paul Werbos, Bernard Widrow
1998 📙	**Pulsed Neural Networks** Edited by Wolfgang Maass, Christopher M. Bishop The MIT Press https://direct.mit.edu/books/edited-volume/2001/Pulsed-Neural-Networks	Most practical applications of artificial neural networks are based on a computational model involving the propagation of continuous variables from one processing unit to the next. In recent years, data from neurobiological experiments have made it increasingly clear that biological neural networks, which communicate through pulses, use the timing of the pulses to transmit information and perform computation. This realization has stimulated significant research on pulsed neural networks, including theoretical analyses and model development, neurobiological modeling, and hardware implementation. This book presents the complete spectrum of current research in pulsed neural networks and includes the most important work from many of the key scientists in the field. Terrence J. Sejnowski's foreword, "Neural Pulse Coding," presents an overview of the topic. The first half of the book consists of longer tutorial articles spanning neurobiology, theory, algorithms, and hardware. The second half contains a larger number of shorter research chapters that present more advanced concepts. The contributors use consistent notation and terminology throughout the book.
1998 📙	**Reinforcement Learning** Richard S. Sutton and Andrew G. Barto. The MIT Press https://mitpress.mit.edu/9780262193986/reinforcement-learning/ (This is the 1st edition; 2nd edition in 2018)	Richard Sutton and Andrew Barto provide a clear and simple account of the key ideas and algorithms of reinforcement learning. Their discussion ranges from the history of the field's intellectual foundations to the most recent developments and applications. Reinforcement learning, one of the most active research areas in artificial intelligence, is a computational approach to learning whereby an agent tries to maximize the total amount of reward it receives when interacting with a complex, uncertain environment. In *Reinforcement Learning*, Richard Sutton and Andrew Barto provide a clear and simple account of the key ideas and algorithms of reinforcement learning. Their discussion ranges from the history of the field's intellectual foundations to the most recent developments and applications. The only necessary mathematical background is familiarity with elementary concepts of probability. The book is divided into three parts. Part I defines the reinforcement learning problem in terms of Markov decision processes. Part II provides basic solution methods: dynamic programming, Monte Carlo methods, and temporal-difference learning. Part III presents a unified view of the solution methods and incorporates artificial neural networks, eligibility traces, and planning; the two final chapters present case studies and consider the future of reinforcement learning.
1998	**Alternative Essences of Intelligence** Rodney A. Brooks, Cynthia Breazeal (Ferrell), Robert Irie, Charles C. Kemp, Matthew Marjanović, Brian Scassellati, Matthew M. Williamson AAAI98 https://aaai.org/papers/00961-aaai98-136-alternative-essences-of-intelligence	🦉 **Presents the philosophy behind "Cog", the humanoid robot of the "Nouvelle AI"** We present a novel methodology for building human-like artificially intelligent systems. We take as a model the only existing systems which are universally accepted as intelligent: humans. We emphasize building intelligent systems which are not masters of a single domain, but, like humans, are adept at performing a variety of complex tasks in the real world. Using evidence from cognitive science and neuroscience, we suggest four alternative essences of intelligence to those held by classical AI. These are the parallel themes of development, social interaction, embodiment, and integration. Following a methodology based on these themes, we have built a physical humanoid robot. In this paper we present our methodology and the insights it affords for facilitating learning, simplifying the computation underlying rich behavior, and building systems that can scale to more complex tasks in more challenging environments
1999 📙	**Cambrian Intelligence: The Early History of the New AI** Rodney A. Brooks MIT Press https://direct.mit.edu/books/monograph/4667/Cambrian-IntelligenceThe-Early-History-of-the-New	Until the mid-1980s, AI researchers assumed that an intelligent system doing high-level reasoning was necessary for the coupling of perception and action. In this traditional model, cognition mediates between perception and plans of action. Realizing that this core AI, as it was known, was illusory, Rodney A. Brooks turned the field of AI on its head by introducing the behavior-based approach to robotics. The cornerstone of behavior-based robotics is the realization that the coupling of perception and action gives rise to all the power of intelligence and that cognition is only in the eye of an observer. Behavior-based robotics has been the basis of successful applications in entertainment, service industries, agriculture, mining, and the home. It has given rise to both autonomous mobile robots and more recent humanoid robots such as Brooks' Cog. This book represents Brooks' initial formulation of and contributions to the development of the behavior-based approach to robotics. It presents all of the key philosophical and technical ideas that put this "bottom-up" approach at the forefront of current research in not only AI but all of cognitive science.
2000-2004		
2000-04	**A Theory of Universal Artificial Intelligence based on Algorithmic Complexity** Marcus Hutter https://arxiv.org/abs/cs/0004001 See also: https://en.wikipedia.org/wiki/AIXI	🦉 **Uses algorithmic complexity theory to define the "most intelligent unbiased agent computable": AIXI model. Of course, it's uncomputable.** Decision theory formally solves the problem of rational agents in uncertain worlds if the true environmental prior probability distribution is known. Solomonoff's theory of universal induction formally solves the problem of sequence prediction for unknown prior distribution. We combine both ideas and get a parameterless theory of universal Artificial Intelligence. We give strong arguments that the resulting AIXI model is the most intelligent unbiased agent possible. We outline for a number of problem classes, including sequence prediction, strategic games, function minimization, reinforcement and supervised learning, how the AIXI model can formally solve them. The major drawback of the AIXI model is that it is uncomputable. To overcome this problem, we construct a modified algorithm AIXI-tl, which is still effectively more intelligent than any other time t and space l bounded agent. The computation time of AIXI-tl is of the order tx2^l. Other discussed topics are formal definitions of intelligence order relations, the horizon problem and relations of the AIXI theory to other AI approaches.
2001-01	**Rapid Object Detection using a Boosted Cascade of Simple Features** Paul Viola, Michael Jones Published in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001 https://ieeexplore.ieee.org/document/990517 https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf	🦉 **Performing image processing without deep NN.** This paper describes a machine learning approach for visual object detection which is capable of processing images extremely rapidly and achieving high detection rates. This work is distinguished by three key contributions. The first is the introduction of a new image representation called the “Integral Image” which allows the features used by our detector to be computed very quickly. The second is a learning algorithm, based on AdaBoost, which selects a small number of critical visual features from a larger set and yields extremely efficient classifiers. The third contribution is a method for combining increasingly more complex classifiers in a “cascade” which allows background regions of the image to be quickly discarded while spending more computation on promising object-like regions. The cascade can be viewed as an object specific focus-of-attention mechanism which unlike previous approaches provides statistical guarantees that discarded regions are unlikely to

		contain the object of interest. In the domain of face detection the system yields detection rates comparable to the best previous systems. Used in real-time applications, the detector runs at 15 frames per second without resorting to image differencing or skin color detection.
2001-08	**Statistical Modeling: The Two Cultures** Leo Breiman Appears in: "Statistical Science" 2001, Vol. 16, No. 3, 199–231 https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full	🦉 **A provocative attack on statistical modeling** There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.
2002 📙📙	**The Handbook of Brain Theory and Neural Networks (Second Edition)** Edited by Michael A. Arbib The MIT Press https://direct.mit.edu/books/edited-volume/4358/The-Handbook-of-Brain-Theory-and-Neural-Networks	Dramatically updating and extending the first edition, published in 1995, the second edition of *The Handbook of Brain Theory and Neural Networks* presents the enormous progress made in recent years in the many subfields related to the two great questions: How does the brain work? and, How can we build intelligent machines? Once again, the heart of the book is a set of almost 300 articles covering the whole spectrum of topics in brain theory and neural networks. The first two parts of the book, prepared by Michael Arbib, are designed to help readers orient themselves in this wealth of material. Part I provides general background on brain modeling and on both biological and artificial neural networks. Part II consists of "Road Maps" to help readers steer through articles in part III on specific topics of interest. The articles in part III are written so as to be accessible to readers of diverse backgrounds. They are cross-referenced and provide lists of pointers to Road Maps, background material, and related reading. The second edition greatly increases the coverage of models of fundamental neurobiology, cognitive neuroscience, and neural network approaches to language. It contains 287 articles, compared to the 266 in the first edition. Articles on topics from the first edition have been updated by the original authors or written anew by new authors, and there are 106 articles on new topics.
2004 📙	**Between Human and Machine: Feedback, Control, and Computing before Cybernetics** David A. Mindell Johns Hopkins Studies in the History of Technology https://press.jhu.edu/books/title/1207/between-human-and-machine	Today, we associate the relationship between feedback, control, and computing with Norbert Wiener's 1948 formulation of cybernetics. But the theoretical and practical foundations for cybernetics, control engineering, and digital computing were laid earlier, between the two world wars. In *Between Human and Machine: Feedback, Control, and Computing before Cybernetics*, David A. Mindell shows how the modern sciences of systems emerged from disparate engineering cultures and their convergence during World War II. Mindell examines four different arenas of control systems research in the United States between the world wars: naval fire control, the Sperry Gyroscope Company, the Bell Telephone Laboratories, and Vannevar Bush's laboratory at MIT. Each of these institutional sites had unique technical problems, organizational imperatives, and working environments, and each fostered a distinct engineering culture. Each also developed technologies to represent the world in a machine. At the beginning of World War II, President Roosevelt established the National Defense Research Committee, one division of which was devoted to control systems. Mindell shows how the NDRC brought together representatives from the four pre-war engineering cultures, and how its projects synthesized conceptions of control, communications, and computing. By the time Wiener articulated his vision, these ideas were already suffusing through engineering. They would profoundly influence the digital world. As a new way to conceptualize the history of computing, this book will be of great interest to historians of science, technology, and culture, as well as computer scientists and theorists. Review: https://web.mit.edu/slava/homepage/articles/Gerovitch%20-%20Review%20of%20Mindell.pdf
2004-03 📅	The **DARPA Grand Challenge** takes place in the Mojave Desert on March 13	• The goal: Develop an autonomous vehicle capable of navigating 142 miles (229 km) of off-road terrain. • Result: No vehicle finished. The best attempt (by Carnegie Mellon's Red Team) only made it 7.4 miles before getting stuck.
2004-11	**Moving AI Out of its Infancy: Changing our Preconceptions** Steve Grand Appears in: "IEEE Intelligent Systems", November/December 2004 https://ieeexplore.ieee.org/document/1363738	Explicit symbolic logic has faded from prominence, but the close coupling of AI and the digital computer, and of thought and the stepwise algorithm, seem about as strong and unquestioned as ever. Of course there's connectionism, but this too is mired in false assumptions that date back a long way. And it seems to have dragged neuroscience down with it to the extent that we now seem unable to think about real brains without resorting to models that owe too much of their inspiration to the three-layer perceptron. Traditional AI has excelled at solving certain kinds of problems. It can make systems that learn but not in any generally applicable way. AI is about making machines do what humans use intelligence to do, and often this doesn't actually require the machines to show any intelligence at all. But for many tasks, especially in robotics, the ability to see, learn, and perform complex motor actions is a prerequisite that the traditional approach has utterly failed to fulfill. • Provocative statement #1: Brains exist to compensate for the slowness of nerves • Provocative statement #2: Brains don't make decisions, but simply reduce the tension between how things are and how we expect or would like them to be. • Provocative statement #3: Brains perform coordinate transforms • Provocative statement #4: Nervous tissue is a new state of matter • Provocative statement #5: The more complex the robot, the easier it is to make progress
2005-2009		
2005-06	**What kind of a graphical model is the brain?** Geoffrey E. Hinton Appears in: "IJCAI'05: Proceedings of the 19th international joint conference on Artificial intelligence"	If neurons are treated as latent variables, our visual systems are non-linear, densely-connected graphical models containing billions of variables and thousands of billions of parameters. Current algorithms would have difficulty learning a graphical model of this scale. Starting with an algorithm that has difficulty learning more than a few thousand parameters, I describe a series of progressively better learning algorithms all of which are designed to run on neuron-like hardware. The latest member of this series can learn deep, multi-layer belief nets quite rapidly. It turns a generic network with three hidden layers and 1.7 million connections into a very

	https://dl.acm.org/doi/10.5555/1642293.1642643 https://www.cs.toronto.edu/~hinton/absps/ijcai05.pdf	good generative model of handwritten digits. After learning, the model gives classification performance that is comparable to the best discriminative methods.
2005-10	The second **DARPA Grand Challenge** takes place in the Mojave Desert on October 8	The goal: navigate a 132-mile (212 km) desert course. Five teams completed the course and Stanford University's "Stanley", an autonomous VW Touareg, finishes in 6 hours, 53 minutes.
2006 📔	**Principles of Interactive Computation** Dina Goldin and Peter Wegner Appears in: "Interactive Computation - The New Paradigm" edited by Dina Goldin, Scott A. Smolka, Peter Wegner https://link.springer.com/book/10.1007/3-540-34874-3	🦉 **How "interactive computation" differs from traditional "algorithms" running in closed universes** This chapter explores the authors' 10-year contributions to interactive computing, with special emphasis on the philosophical question of how truth has been used and misused in computing and other disciplines. We explore the role of rationalism and empiricism in formulating true principles of computer science, politics, and religion. We show that interaction is an empiricist rather than rationalist principle, and that rationalist proponents of computing have been the strongest opponents of our belief that interaction provides an empirical foundation for both computer problem solving and human behavior. The rationalist position was adopted by Pythagoras, Descartes, Kant, and many modern philosophers; our interactive approach to computing suggests that empiricism provides a better framework for understanding principles of computing.
2006	**Hierarchical Temporal Memory Concepts, Theory, and Terminology** Jeff Hawkins, Dileep George https://dileeplearning.github.io/uploads/HTM.pdf	🦉 **A cognitive theory** HTMs are unlike traditional programmable computers. With traditional computers, a programmer creates specific programs to solve specific problems. For example, one program may be used to recognize speech and another completely different program may be used to model weather. HTM, on the other hand, is best thought of as a memory system. HTMs are not programmed and do not execute different algorithms for different problems. Instead, HTMs "learn" how to solve problems. HTMs are trained by exposing them to sensory data and the capability of the HTM is determined largely by what it has been exposed to.
2006-05	**Mindless Intelligence** Jordan B. Pollack, Brandeis University Appears in: "IEEE Intelligent Systems", May/June 2006 https://ieeexplore.ieee.org/document/1637350 https://ia601000.us.archive.org/33/items/pollackmindlessintelligence/pollack%20mindless%20intelligence.pdf	AI has stalled because of its preoccupation with simulating the human mind. By studying intelligence in natural systems, outside the mind, we can reinvigorate the field
2007-02	**Antipatterns in the Creation of Intelligent Systems** Phil Laplante, Robert R. Hoffman, Gary Klein Appears in: "IEEE Intelligent Systems", January/February 2007 "Human-Centered Computing" Column https://ieeexplore.ieee.org/document/4078960 https://www.ihmc.us/wp-content/uploads/2021/04/23.-Antipatterns.pdf	A design pattern is a named problem-solution pair that enables large-scale reuse of software architectures or their components. Ideally, patterns explicitly capture expert knowledge, design trade-offs, and design rationale and make these lessons learned widely available for off-the-shelf use. They can also enhance developers' vocabulary - for example, by easing the transition to object-oriented programming. Conventionally, patterns consist of four elements: a name, the problem to be solved, the solution to the problem (often termed the refactored solution), and the consequences of the solution. Numerous sets of patterns (collectively known as pattern languages) exist for software design, analysis, management, and so on. Shortly after the notion of design patterns emerged, practitioners began discussing problem-solution pairs in which the solution did more harm than good. These have come to be known as antipatterns, and they are well known in the design and management communities
2007-05	**In the News: Bridging the Gap between Neuroscience and AI** Sara Reese Hedberg Appears: "IEEE Intelligent Systems", May/June 2007 https://ieeexplore.ieee.org/document/4216972	The human brain is a three-pound, gnarled lump of 100 billion neurons and the wondrous control center of the human body and intelligence. Even with today's technology, we have only limited understanding of how it works and how to embody some of its intelligence in machines.

2007-06 📔	**Artificial General Intelligence**
	Edited by Ben Goertzel, Cassio Pennachin
	Springer
	https://link.springer.com/book/10.1007/978-3-540-68677-4
		There is actually nothing particular remarkable about this book except that I have it somewhere in the attic, and its price: from the already rather high EUR 67 in 2007 (hardcover), it is now (early 2025) available at an eye-watering EUR 246 (also hardcover). 🧐
		• Contemporary Approaches to Artificial General Intelligence (Cassio Pennachin, Ben Goertzel)
		• The Logic of Intelligence (Pei Wang)
		    • https://cis.temple.edu/~pwang/Publication/logic_intelligence.pdf
		• The Novamente Artificial Intelligence Engine (Ben Goertzel, Cassio Pennachin)
		• Essentials of General Intelligence: The Direct Path to Artificial General Intelligence (Peter Voss)
		• Artificial Brains (Hugo de Garis) (this is about the CAM Brain Machine)
		• The New AI: General & Sound & Relevant for Physics (Jürgen Schmidhuber)
		    • https://arxiv.org/pdf/cs/0302012
		• Gödel Machines: Fully Self-referential Optimal Universal Self-improvers (Jürgen Schmidhuber)
		    • https://arxiv.org/abs/cs/0309048
		• Universal Algorithmic Intelligence: A Mathematical Top→Down Approach (Marcus Hutter)
		    • https://arxiv.org/abs/cs/0701125
		• Program Search as a Path to Artificial General Intelligence (Lukasz Kaiser)
		• The Natural Way to Artificial Intelligence (Vladimir G. Red'ko)
		• 3D Simulation: the Key to A.I. (Keith A. Hoyes)
		• Levels of Organization in General Intelligence (Eliezer Yudkowsky)

2007-11 📆	The **DARPA Urban Challenge** takes place from November 3 to November 6, 2007 at the former George Air Force Base in Victorville, California,	• The goal: navigate a city-like environment, follow traffic laws, merge, and avoid obstacles. • Winner: Carnegie Mellon's "Boss", a modified Chevy Tahoe, completes the challenge in **6 hours and 53 minutes**, demonstrating significant progress in autonomous driving technologies.
2007-12 📘	**Top 10 algorithms in data mining** Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh et al. Top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006. https://link.springer.com/article/10.1007/s10115-007-0114-2 (Paywalled but widely distributed)	Simple approaches in data mining and machine learning may sometimes be sufficient. Listed: • C4.5 (trees) • k-means • Support Vector Machines • Apriori • Expectation Maximization (not handled in this book) • PageRank (not handled in this book) • AdaBoost • k-Nearest Neighbors • Naïve Bayes • CART
2009-08 📘	**Spiking Neural Networks: Review Article** Samanwoy Ghosh-Dastidar, Hojjat Adeli Appears in: "International Journal of Neural Systems", Vol. 19, No. 04 https://worldscientific.com/doi/10.1142/S0129065709002002 (Paywalled, what is an open equivalent?)	Most current Artificial Neural Network (ANN) models are based on highly simplified brain dynamics. They have been used as powerful computational tools to solve complex pattern recognition, function estimation, and classification problems. ANNs have been evolving towards more powerful and more biologically realistic models. In the past decade, Spiking Neural Networks (SNNs) have been developed which comprise of *spiking neurons*. Information transfer in these neurons mimics the information transfer in biological neurons, i.e., via the precise timing of spikes or a sequence of spikes. To facilitate learning in such networks, new learning algorithms based on varying degrees of biological plausibility have also been developed recently. Addition of the temporal dimension for information encoding in SNNs yields new insight into the dynamics of the human brain and could result in compact representations of large neural networks. As such, SNNs have great potential for solving complicated time-dependent pattern recognition problems because of their inherent dynamic representation. This article presents a state-of-the-art review of the development of spiking neurons and SNNs, and provides insight into their evolution as the third generation neural networks.
2010-2014		
2010-01	**Understanding the difficulty of training deep feedforward neural networks** Xavier Glorot, Yoshua Bengio Appears in: "Proceedings of the 13th International Conference on Artificial Intelligence and Statistics" https://proceedings.mlr.press/v9/glorot10a/glorot10a.pdf	🦉 **Introduction of Xavier Initialization and ReLU Activation Function to handle exploding/vanishing gradient** Whereas before 2006 it appears that deep multilayer neural networks were not successfully trained, since then several algorithms have been shown to successfully train them, with experimental results showing the superiority of deeper vs less deep architectures. All these experimental results were obtained with new initialization or training mechanisms. Our objective here is to understand better why standard gradient descent from random initialization is doing so poorly with deep neural networks, to better understand these recent relative successes and help design better algorithms in the future. We first observe the influence of the non-linear activations function. We find that the logistic sigmoid activation is unsuited for deep networks with random initialization because of its mean value, which can drive especially the top hidden layer into saturation. Surprisingly, we find that saturated units can move out of saturation by themselves, albeit slowly, and explaining the plateaus sometimes seen when training neural networks. We find that a new non-linearity that saturates less can often be beneficial. Finally, we study how activations and gradients vary across layers and during training, with the idea that training may be more difficult when the singular values of the Jacobian associated with each layer are far from 1. Based on these considerations, we propose a new initialization scheme that brings substantially faster convergence
2010-08	**Building Watson: An Overview of the DeepQA Project** David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty https://research.ibm.com/publications/building-watson-an-overview-of-the-deepqa-project Appears in: "AI Magazine" (The Magazine of AAAI), Fall 2010	🦉 **A great attempt at building a Question Answering System before the advent of LLMs (it does not hallucinate)** IBM Research undertook a challenge to build a computer system that could compete at the human champion level in real time on the American TV quiz show, Jeopardy. The extent of the challenge includes fielding a real-time automatic contestant on the show, not merely a laboratory exercise. The Jeopardy Challenge helped us address requirements that led to the design of the DeepQA architecture and the implementation of Watson. After three years of intense research and development by a core team of about 20 researchers, Watson is performing at human expert levels in terms of precision, confidence, and speed at the Jeopardy quiz show. Our results strongly suggest that DeepQA is an effective and extensible architecture that can be used as a foundation for combining, deploying, evaluating, and advancing a wide range of algorithmic techniques to rapidly advance the field of question answering (QA).
2012-06	**An overview of Hierarchical Temporal Memory: A new neocortex algorithm** Xi Chen; Wei Wang; Wei Li Appears in: "2012 Proceedings of International Conference on Modelling, Identification and Control", 24-26 June 2012 https://ieeexplore.ieee.org/abstract/document/6260285	The overview presents the development and application of Hierarchical Temporal Memory (HTM). HTM is a new machine learning method which was proposed by Jeff Hawkins in 2005. It is a biologically inspired cognitive method based on the principle of how human brain works. The method invites hierarchical structure and proposes a memory-prediction framework, thus making it able to predict what will happen in the near future. This overview mainly introduces the developing process of HTM, as well as its principle, characteristics, advantages and applications in vision, image processing and robots movement, some potential applications by using HTM, such as thinking process, are also put forward.
2013	**Essentials of Metaheuristics, 2nd Edition** Sean Luke http://cs.gmu.edu/~sean/book/metaheuristics/	🦉 **Overview work of optimization algorithms** • Gradient-based Optimization • Single-State Methods (Hill Climbing, Single-State Global Optimization Algorithms, Adjusting the Modification Procedure, Simulated Annealing, Tabu Search, Iterated Local Search) • Population Methods (Evolution Strategies, The Genetic Algorithm, Exploitative Variations, Differential Evolution, Particle Swarm Optimization) • Representation (Vectors, Direct Encoded Graphs, Trees and Genetic Programming, Lists, Rulesets, Bloat)

		• Parallel Methods (Multiple Threads, Island Models, Master-Slave Fitness Assessment, Spatially Embedded Models) • Coevolution (1-Population Competitive Coevolution, 2-Population Competitive Coevolution, N-Population Cooperative Coevolution, Niching) • Multiobjective Optimization (Naive Methods, Non-Dominated Sorting, Pareto Strength) • Combinatorial Optimization (General-Purpose Optimization and Hard Constraints, Ant Colony Optimization, Guided Local Search) • Optimization by Model Fitting (Model Fitting by Classification, Model Fitting with a Distribution) • Policy Optimization (Reinforcement Learning: Dense Policy Optimization, Sparse Stochastic Policy Optimization, Pitt Approach Rule System, Michigan Approach Learning Classifier Systems, Regression with the Michigan Approach, Is this Genetic Programming?) • Miscellany (Experimental Methodology, Simple Test Propbelms, Where to Go Next, Example Course Syllabi for the Text)
2013-05	**The Seven Deadly Myths of "Autonomous Systems"** Jeffrey M. Bradshaw, Robert R. Hoffman, Matthew Johnson, and David D. Woods Appears in "IEEE Intelligent Systems", May/June 2013, Human-Centered Computing Column	• Myth 1: "Autonomy" is unidimensional. • Myth 2: The conceptualization of "levels of autonomy" is a useful scientific grounding for the development of autonomous system roadmaps. • Myth 3: Autonomy is a widget. • Myth 4: Autonomous systems are autonomous. • Myth 5: Once achieved, full autonomy obviates the need for human-machine collaboration. • Myth 6: As machines acquire more autonomy, they will work as simple substitutes (or multipliers) of human capability. • Myth 7: "Full autonomy" is not only possible, but is always desirable.
2013-12	**Playing Atari with Deep Reinforcement Learning** Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, Martin Riedmiller https://arxiv.org/abs/1312.5602 *And also, one year later:* **Human-level control through deep reinforcement learning** Volodymyr Mnih, Koray Kavukcuoglu, David Silver et al. Letter to "Nature", Nature volume 518, pages 529–533 (2015-02-26) https://www.nature.com/articles/nature14236 (paywalled)	🦉 **First major success of Deep Reinforcement Learning** We present the first deep learning model to successfully learn control policies directly from high-dimensional sensory input using reinforcement learning. The model is a convolutional neural network, trained with a variant of Q-learning, whose input is raw pixels and whose output is a value function estimating future rewards. We apply our method to seven Atari 2600 games from the Arcade Learning Environment, with no adjustment of the architecture or learning algorithm. We find that it outperforms all previous approaches on six of the games and surpasses a human expert on three of them.
2014-01	**Dropout: a simple way to prevent neural networks from overfitting** Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov Appears in: "The Journal of Machine Learning Research, Volume 15, Issue 1" https://dl.acm.org/doi/10.5555/2627435.2670313	🦉 **Introducing Dropout** Deep neural nets with a large number of parameters are very powerful machine learning systems. However, overfitting is a serious problem in such networks. Large networks are also slow to use, making it difficult to deal with overfitting by combining the predictions of many different large neural nets at test time. Dropout is a technique for addressing this problem. The key idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much. During training, dropout samples from an exponential number of different "thinned" networks. At test time, it is easy to approximate the effect of averaging the predictions of all these thinned networks by simply using a single unthinned network that has smaller weights. This significantly reduces overfitting and gives major improvements over other regularization methods. We show that dropout improves the performance of neural networks on supervised learning tasks in vision, speech recognition, document classification and computational biology, obtaining state-of-the-art results on many benchmark data sets.
2014-09	**Very Deep Convolutional Networks for Large-Scale Image Recognition** Karen Simonyan, Andrew Zisserman https://arxiv.org/abs/1409.1556 Latest version 10 Apr 2015.	🦉 **Demonstrates deep CNNs, Visual Geometry Group (VGG) network, and model pretraining** In this work we investigate the effect of the convolutional network depth on its accuracy in the large-scale image recognition setting. Our main contribution is a thorough evaluation of networks of increasing depth using an architecture with very small (3x3) convolution filters, which shows that a significant improvement on the prior-art configurations can be achieved by pushing the depth to 16-19 weight layers. These findings were the basis of our ImageNet Challenge 2014 submission, where our team secured the first and the second places in the localisation and classification tracks respectively. We also show that our representations generalise well to other datasets, where they achieve state-of-the-art results. We have made our two best-performing ConvNet models publicly available to facilitate further research on the use of deep visual representations in computer vision.
2014-09	**Neural Machine Translation by Jointly Learning to Align and Translate** Dzmitry Bahdanau, KyungHyun Cho, Yoshua Bengio Latest version is v7 of 2016-05-19 https://arxiv.org/abs/1409.0473	🦉 **Introduction of the "attention mechanism" for neural language processing and neural machine translation, the foundation for later Transformer Models** Neural machine translation is a recently proposed approach to machine translation. Unlike the traditional statistical machine translation, the neural machine translation aims at building a single neural network that can be jointly tuned to maximize the translation performance. The models proposed recently for neural machine translation often belong to a family of encoder-decoders and consists of an encoder that encodes a source sentence into a fixed-length vector from which a decoder generates a translation. In this paper, we conjecture that the use of a fixed-length vector is a bottleneck in improving the performance of this basic encoder-decoder architecture, and propose to extend this by allowing a model to automatically (soft-)search for parts of a source sentence that are relevant to predicting a target word, without having to form these parts as a hard segment explicitly. With this new approach, we achieve a translation performance comparable to the existing state-of-the-art phrase-based system on the task of English-to-French translation. Furthermore, qualitative analysis reveals that the (soft-)alignments found by the model agree well with our intuition.
2014-11	**Machine Learning: The High-Interest Credit Card of Technical Debt** D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young https://research.google.com/pubs/pub43146.html	Machine learning offers a fantastically powerful toolkit for building complex systems quickly. This paper argues that it is dangerous to think of these quick wins as coming for free. Using the framework of technical debt, we note that it is remarkably easy to incur massive ongoing maintenance costs at the system level when applying machine learning. The goal of this paper is highlight several machine learning specific risk factors and design patterns to be avoided or refactored where possible. These include boundary erosion, entanglement, hidden feedback loops, undeclared consumers, data dependencies, changes in the external world, and a variety of system-level anti-patterns.

2014-11	**Seven Cardinal Virtues of Human-Machine Teamwork: Examples from the DARPA Robotic Challenge** Matthew Johnson, Jeffrey M. Bradshaw, Robert R. Hoffman et. al. Appears in: "IEEE Intelligent Systems", November/December 2014, Human-Centered Computing Column Paywalled: https://ieeexplore.ieee.org/document/6982119 Open: https://www.researchgate.net/publication/273393961_Seven_Cardinal_Virtues_of_Human-Machine_Teamwork_Examples_from_the_DARPA_Robotic_Challenge	1. Focus on improving mission performance of the work system, not on maximizing autonomous capabilities. 2. Assess the sweet spot in development effort payoff. 3. If you don't plan to fail, you fail to plan. 4. Think "combine and succeed," not "divide and conquer." 5. Design for teamwork in addition to taskwork. 6. Designing for human-machine teamwork goes deeper than the user interface. 7. Don't simply downsize human involvement; rightsize it.
2014-12	**Adam: A Method for Stochastic Optimization** Diederik P. Kingma, Jimmy Lei Ba https://arxiv.org/abs/1412.6980 Latest version: 2017-01	🦉 **Introduces the "Adam" algorithm** We introduce Adam, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments. The method is straightforward to implement, is computationally efficient, has little memory requirements, is invariant to diagonal rescaling of the gradients, and is well suited for problems that are large in terms of data and/or parameters. The method is also appropriate for non-stationary objectives and problems with very noisy and/or sparse gradients. The hyper-parameters have intuitive interpretations and typically require little tuning. Some connections to related algorithms, on which Adam was inspired, are discussed. We also analyze the theoretical convergence properties of the algorithm and provide a regret bound on the convergence rate that is comparable to the best known results under the online convex optimization framework. Empirical results demonstrate that Adam works well in practice and compares favorably to other stochastic optimization methods. Finally, we discuss AdaMax, a variant of Adam based on the infinity norm.
2014-12	**A Common Logic to Seeing Cats and Cosmos** Popular Science explainer at Quanta Magazine by Natalie Wolchover https://www.quantamagazine.org/a-common-logic-to-seeing-cats-and-cosmos-20141204/	New research suggests physicists, computers and brains employ the same procedure to tease out important features from among other irrelevant bits of data.
2015-2016		
2015 📙	**Regression Modeling Strategies** With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis Second Edition Springer Frank E. Harrell, Jr.	
2015 📙	**Our Robots, Ourselves** Robotics and the Myths of Autonomy David A. Mindell	Book presentation: https://www.youtube.com/watch?v=4nDdqGUMdAY
2015-06 📆	The final event of the DARPA Robotics Challenge takes place in June 2015 at Fairplex in Pomona, California.	Goal: advance robotic systems for disaster response scenarios. Teams had to navigate a simulated disaster environment where the robot would perform tasks like: • Driving a vehicle • Opening doors • Climbing stairs • Using tools like a drill • Removing debris Team TRI (Toyota Engineering Society), the team behind "CHIMP" (a robot developed by Carnegie Mellon University, the "CMU Highly Intelligent Mobile Platform") wins.
2015-06	**Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification** Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun https://arxiv.org/abs/1502.01852	🦉 **Introduction of He Initialization** Rectified activation units (rectifiers) are essential for state-of-the-art neural networks. In this work, we study rectifier neural networks for image classification from two aspects. First, we propose a Parametric Rectified Linear Unit (PReLU) that generalizes the traditional rectified unit. PReLU improves model fitting with nearly zero extra computational cost and little overfitting risk. Second, we derive a robust initialization method that particularly considers the rectifier nonlinearities. This method enables us to train extremely deep rectified models directly from scratch and to investigate deeper or wider network architectures. Based on our PReLU networks (PReLU-nets), we achieve 4.94% top-5 test error on the ImageNet 2012 classification dataset. This is a 26% relative improvement over the ILSVRC 2014 winner (GoogLeNet, 6.66%). To our knowledge, our result is the first to surpass human-level performance (5.1%, Russakovsky et al.) on this visual recognition challenge.
2015-17	**Gradient Estimation Using Stochastic Computation Graphs** John Schulman, Nicolas Heess, Theophane Weber, Pieter Abbeel https://arxiv.org/abs/1506.05254	In a variety of problems originating in supervised, unsupervised, and reinforcement learning, the loss function is defined by an expectation over a collection of random variables, which might be part of a probabilistic model or the external world. Estimating the gradient of this loss function, using samples, lies at the core of gradient-based learning algorithms for these problems. We introduce the formalism of stochastic computation graphs—directed acyclic graphs that include both deterministic functions and conditional probability distributions—and describe how to easily and automatically derive an unbiased estimator of the loss function's gradient. The

	Latest version: 2016-01-05	resulting algorithm for computing the gradient estimator is a simple modification of the standard backpropagation algorithm. The generic scheme we propose unifies estimators derived in variety of prior work, along with variance-reduction techniques therein. It could assist researchers in developing intricate models involving a combination of stochastic and deterministic operations, enabling, for example, attention, memory, and control actions.
2015-09	**Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence** Gary Marcus, Ernest Davis Appears in: "Communications of the ACM", September 2015 https://dl.acm.org/doi/10.1145/2701413	AI has seen great advances of many kinds recently, but there is one critical area where progress has been extremely slow: ordinary commonsense.
2015-12	**Deep Residual Learning for Image Recognition** Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun (Microsoft Research) CVPR 2016 (IEEE Conference on Computer Vision and Pattern Recognition) https://arxiv.org/abs/1512.03385	Introduces residual connections (skip connections), allowing gradients to flow directly through shortcut paths during backpropagation. Mitigates the vanishing gradient problem and enables the training of extremely deep networks (e.g., ResNet-50, ResNet-101). Demonstrates a significant improvement in image classification tasks, winning the ImageNet competition in 2015.
2016-03	**Is AlphaGo Really Such a Big Deal?** Popular Science explainer at Quanta Magazine by Michael Nielsen https://www.quantamagazine.org/is-alphago-really-such-a-big-deal-20160329/	"The Go-playing program captures elements of human intuition, an advance that promises far-reaching consequences."
2016-03	**Why Neurons Have Thousands of Synapses, a Theory of Sequence Memory in Neocortex** Jeff Hawkins, Subutai Ahmad Appears in: "Front Neural Circuits", March 2016 https://pubmed.ncbi.nlm.nih.gov/27065813/	Pyramidal neurons represent the majority of excitatory neurons in the neocortex. Each pyramidal neuron receives input from thousands of excitatory synapses that are segregated onto dendritic branches. The dendrites themselves are segregated into apical, basal, and proximal integration zones, which have different properties. It is a mystery how pyramidal neurons integrate the input from thousands of synapses, what role the different dendrites play in this integration, and what kind of network behavior this enables in cortical tissue. It has been previously proposed that non-linear properties of dendrites enable cortical neurons to recognize multiple independent patterns. In this paper we extend this idea in multiple ways. First we show that a neuron with several thousand synapses segregated on active dendrites can recognize hundreds of independent patterns of cellular activity even in the presence of large amounts of noise and pattern variation. We then propose a neuron model where patterns detected on proximal dendrites lead to action potentials, defining the classic receptive field of the neuron, and patterns detected on basal and apical dendrites act as predictions by slightly depolarizing the neuron without generating an action potential. By this mechanism, a neuron can predict its activation in hundreds of independent contexts. We then present a network model based on neurons with these properties that learns time-based sequences. The network relies on fast local inhibition to preferentially activate neurons that are slightly depolarized. Through simulation we show that the network scales well and operates robustly over a wide range of parameters as long as the network uses a sparse distributed code of cellular activations. We contrast the properties of the new network model with several other neural network models to illustrate the relative capabilities of each. We conclude that pyramidal neurons with thousands of synapses, active dendrites, and multiple integration zones create a robust and powerful sequence memory. Given the prevalence and similarity of excitatory neurons throughout the neocortex and the importance of sequence memory in inference and behavior, we propose that this form of sequence memory may be a universal property of neocortical tissue.
2016-08	**Densely Connected Convolutional Networks** Gao Huang, Zhuang Liu, Laurens van der Maaten, Kilian Q. Weinberger CVPR 2017 (IEEE Conference on Computer Vision and Pattern Recognition) https://arxiv.org/abs/1608.06993 (latest version 2018-01-28)	Recent work has shown that convolutional networks can be substantially deeper, more accurate, and efficient to train if they contain shorter connections between layers close to the input and those close to the output. In this paper, we embrace this observation and introduce the Dense Convolutional Network (DenseNet), which connects each layer to every other layer in a feed-forward fashion. Whereas traditional convolutional networks with L layers have L connections - one between each layer and its subsequent layer - our network has L(L+1)/2 direct connections. For each layer, the feature-maps of all preceding layers are used as inputs, and its own feature-maps are used as inputs into all subsequent layers. DenseNets have several compelling advantages: they alleviate the vanishing-gradient problem, strengthen feature propagation, encourage feature reuse, and substantially reduce the number of parameters. We evaluate our proposed architecture on four highly competitive object recognition benchmark tasks (CIFAR-10, CIFAR-100, SVHN, and ImageNet). DenseNets obtain significant improvements over the state-of-the-art on most of them, whilst requiring less computation to achieve high performance.
2016-09	**Bridging the Ethical Gap: From Human Principles to Robot Instructions** Neil McBride, Robert R. Hoffman Appears in "IEEE Intelligent Systems", September/October 2016, Human-Centered Computing Column https://ieeexplore.ieee.org/document/7579396 https://www.ihmc.us/wp-content/uploads/2021/04/64.-Robotics-Ethical-Gap.pdf	Asimov's three laws of robotics and the Murphy-Woods alternative laws assume that a robot has the cognitive ability to make moral decisions, and fail to escape the myth of self-sufficiency. But ethical decision making on the part of robots in human-robot interaction is grounded on the interdependence of human and machine. Furthermore, the proposed laws are high-level principles that cannot easily be translated into machine instructions because there is an immense gap between the architecture, implementation, and activity of humans and robots in addressing ethical situations. The characterization of the ethical gap, particularly with reference to the Murphy-Woods laws, leads to a proposal for a shift in focus away from the autonomous behavior of the robot to human-robot communication at the interface, and the development of interdependence rules to underpin the process of ethical decision-making.
2017		
2017-01 📘	**Deep Reinforcement Learning: An Overview** Yuxi Li https://arxiv.org/abs/1701.07274 Latest edition: 26 Nov 2018 *Significant update:* **Deep Reinforcement Learning**, 15 October 2018 https://arxiv.org/abs/1810.06339	We give an overview of recent exciting achievements of deep reinforcement learning (RL). We discuss six core elements, six important mechanisms, and twelve applications. We start with background of machine learning, deep learning and reinforcement learning. Next we discuss core RL elements, including value function, in particular, Deep Q-Network (DQN), policy, reward, model, planning, and exploration. After that, we discuss important mechanisms for RL, including attention and memory, unsupervised learning, transfer learning, multi-agent RL, hierarchical RL, and learning to learn. Then we discuss various applications of RL, including games, in particular, AlphaGo, robotics, natural language processing, including dialogue systems, machine translation, and text generation, computer vision, neural architecture design, business management, finance, healthcare, Industry 4.0, smart grid, intelligent transportation systems, and computer systems. We mention topics not reviewed yet, and list a collection of RL resources. After presenting a brief summary, we close with discussions.
2017-02 📘	**A Survey of Learning Classifier Systems in Games [Review Article]**	🦉 **An approach at eliciting condition-action rules through genetic algorithms and RL (not NN)**

	Kamran Shafi; Hussein A. Abbass Appears in: "IEEE Computational Intelligence Magazine", February 2017 https://ieeexplore.ieee.org/document/7807383 (paywalled) http://husseinabbass.net/papers/LCSinGames_Final.pdf (seems to be a preversion)	Games are becoming increasingly indispensable, not only for fun but also to support tasks that are more serious, such as education, strategic planning, and understanding of complex phenomena. Computational intelligence-based methods are contributing significantly to this development. Learning Classifier Systems (LCS) is a pioneering computational intelligence approach that combines machine learning methods with evolutionary computation, to learn problem solutions in the form of interpretable rules. These systems offer several advantages for game applications, including a powerful and flexible agent architecture built on a knowledge-based symbolic modeling engine; modeling flexibility that allows integrating domain knowledge and different machine learning mechanisms under a single computational framework; an ability to adapt to diverse game requirements; and an ability to learn and generate creative agent behaviors in real-time dynamic environments. We present a comprehensive and dedicated survey of LCS in computer games. The survey highlights the versatility and advantages of these systems by reviewing their application in a variety of games. The survey is organized according to a general game classification and provides an opportunity to bring this important research direction into the public eye. We discuss the strengths and weaknesses of the existing approaches and provide insights into important future research directions. See alos: https://en.wikipedia.org/wiki/Learning_classifier_system
2017-03	**Evolution Strategies as a Scalable Alternative to Reinforcement Learning** Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, Ilya Sutskever (OpenAI) https://arxiv.org/abs/1703.03864 Latest version: 2017-09-07	We explore the use of Evolution Strategies (ES), a class of black box optimization algorithms, as an alternative to popular MDP-based RL techniques such as Q-learning and Policy Gradients. Experiments on MuJoCo and Atari show that ES is a viable solution strategy that scales extremely well with the number of CPUs available: By using a novel communication strategy based on common random numbers, our ES implementation only needs to communicate scalars, making it possible to scale to over a thousand parallel workers. This allows us to solve 3D humanoid walking in 10 minutes and obtain competitive results on most Atari games after one hour of training. In addition, we highlight several advantages of ES as a black box optimization technique: it is invariant to action frequency and delayed rewards, tolerant of extremely long horizons, and does not need temporal discounting or value function approximation.
2017-05	**ImageNet classification with deep convolutional neural networks** Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton https://dl.acm.org/doi/10.1145/3065386	🦉 **First image-recognition program to use GPUs** We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0%, respectively, which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully connected layers we employed a recently developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.
2017-06	**Attention Is All You Need** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin https://arxiv.org/abs/1706.03762 Latest version: 2023-08-02	🦉 **Introduction of the "Transformer" architecture** The dominant sequence transduction models are based on complex recurrent or convolutional neural networks in an encoder-decoder configuration. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.
2017-07	**Human-Level Intelligence or Animal-Like Abilities?** Adnan Darwiche https://arxiv.org/abs/1707.04327 Appears in: "Communications of the ACM", October 2018 https://cacm.acm.org/research/human-level-intelligence-or-animal-like-abilities/	**A critical look on the abilities of the new AI agents** "The vision systems of the eagle and the snake outperform everything that we can make in the laboratory, but snakes and eagles cannot build an eyeglass or a telescope or a microscope." (Judea Pearl)
2017-07	**Interactive Cognitive Systems and Social Intelligence** Pat Langley, Institute for the Stufy of Learning and Expertise Appears in: "IEEE Intelligent Systems", Juyl/August 2017 https://ieeexplore.ieee.org/document/8012318 paywalled http://www.isle.org/~langley/papers/interact.ieee17.pdf	Research on cognitive systems adopts the aims and assumptions of classical AI research, emphasizing the construction of intelligent agents that exhibit complex behavior. This article reviews the cognitive systems paradigm and two widely adopted hypotheses-physical symbol systems and heuristic search-that underpin it. The author introduces a third claim-the social cognition hypothesis-that states intelligence requires the ability to represent and reason about others' mental states. The article also examines a number of computational artifacts, both historical and recent, that focus on interaction and exhibit this capacity. Examples include dialogue systems, synthetic experts, believable agents, intelligent tutors, interactive robots, and instructable game players. In closing, the author identifies issues in social cognition that deserve greater attention and poses challenges that can drive future research on interactive cognitive systems.
2017-07	**Computers Play Chess Computers Play Go - Humans Play Dungeons & Dragons** Simon Ellis and James Hendler Appears in: "IEEE Intelligent Systems" July/August 2017 https://ieeexplore.ieee.org/abstract/document/8012338 paywalled	With the AlphaGo program recently beating an expert human player, AI researchers are exploring new challenges, the biggest being human-oriented games such as Dungeons & Dragons.
2017-09 🐿️	**New Theory Cracks Open the Black Box of Deep Learning** Popular Science explainer at Quanta Magazine by Natalie Wolchover https://www.quantamagazine.org/new-theory-cracks-open-the-black-box-of-deep-	"A new idea called the 'information bottleneck' is helping to explain the puzzling success of today’s artificial-intelligence algorithms — and might also explain how human brains learn."

	learning-20170921/	
2017-09 🐿	**Clever Machines Learn How to Be Curious** Popular Science explainer at Quanta Magazine by John Pavlus https://www.quantamagazine.org/clever-machines-learn-how-to-be-curious-20170919/	"Computer scientists are finding ways to code curiosity into intelligent machines."
2017-09 📘	**A Brief Survey of Deep Reinforcement Learning** Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, Anil Anthony Bharath https://arxiv.org/abs/1708.05866	Deep reinforcement learning is poised to revolutionise the field of AI and represents a step towards building autonomous systems with a higher level understanding of the visual world. Currently, deep learning is enabling reinforcement learning to scale to problems that were previously intractable, such as learning to play video games directly from pixels. Deep reinforcement learning algorithms are also applied to robotics, allowing control policies for robots to be learned directly from camera inputs in the real world. In this survey, we begin with an introduction to the general field of reinforcement learning, then progress to the main streams of value-based and policy-based methods. Our survey will cover central algorithms in deep reinforcement learning, including the deep Q-network, trust region policy optimisation, and asynchronous advantage actor-critic. In parallel, we highlight the unique advantages of deep neural networks, focusing on visual understanding via reinforcement learning. To conclude, we describe several current areas of research within the field.
2017-10 🐿	**Artificial Intelligence Learns to Learn Entirely on Its Own** Popular Science explainer at Quanta Magazine by Kevin Hartnett https://www.quantamagazine.org/artificial-intelligence-learns-to-learn-entirely-on-its-own-20171018/	"A new version of AlphaGo needed no human instruction to figure out how to clobber the best Go player in the world — itself."
2018		
2018 📙	**Architects of Intelligence** "The truth about AI from the people building it." Martin Ford PACKT publishing, 2018	Martin Ford ("New York Times Bestselling Author of Rise of the Robots") interviews: Yoshua Bengio, Stuart Russell, Geoffrey Hinton, Nick Bostrom, Yann LeCun, Fei-Fei Li, Demis Hassabis, Andrew Ng, Rana el Kaliouby, Ray Kurzweil, Daniela Rus, James Manyika, Gary Marcus, Barbara Grosz, Judea Pearl, Jeff Dean, Daphne Koller, David Ferruci, Rodney Brooks, Cynthia Breazeal, Josh Tenenbaum, Oren Etzioni, Bryan Johnson
2018-01	**Deep Learning: A Critical Appraisal** Gary Marcus https://arxiv.org/abs/1801.00631	Although deep learning has historical roots going back decades, neither the term “deep learning” nor the approach was popular just over five years ago, when the field was reignited by papers such as Krizhevsky, Sutskever and Hinton’s now classic 2012 (Krizhevsky, Sutskever, & Hinton, 2012) deep net model of Imagenet. What has the field discovered in the five subsequent years? Against a background of considerable progress in areas such as speech recognition, image recognition, and game playing, and considerable enthusiasm in the popular press, I present ten concerns for deep learning, and suggest that deep learning must be supplemented by other techniques if we are to reach artificial general intelligence.
2018-01	**Innateness, AlphaZero and Artificial Intelligence** Gary Marcus https://arxiv.org/abs/1801.05667	The concept of innateness is rarely discussed in the context of artificial intelligence. When it is discussed, or hinted at, it is often the context of trying to reduce the amount of innate machinery in a given system. In this paper, I consider as a test case a recent series of papers by Silver et al (Silver et al., 2017a) on AlphaGo and its successors that have been presented as an argument that a “even in the most challenging of domains: it is possible to train to superhuman level, without human examples or guidance”, “starting tabula rasa.” I argue that these claims are overstated, for multiple reasons. I close by arguing that artificial intelligence needs greater attention to innateness, and I point to some proposals about what that innateness might look like.
2018-01	**Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution** Judea Pearl https://arxiv.org/abs/1801.04016	Machines need to be able to model causality before progress can be made
2018-01	**Superhuman AI for heads-up no-limit poker: Libratus beats top professionals** Noam Brown and Tuomas Sandholm Appears in: Science, 26th January 2018 https://www.science.org/doi/10.1126/science.aao1733	No-limit Texas hold’em is the most popular form of poker. Despite artificial intelligence (AI) successes in perfect-information games, the private information and massive game tree have made no-limit poker difficult to tackle. We present Libratus, an AI that, in a 120,000-hand competition, defeated four top human specialist professionals in heads-up no-limit Texas hold’em, the leading benchmark and long-standing challenge problem in imperfect-information game solving. Our game-theoretic approach features application-independent techniques: an algorithm for computing a blueprint for the overall strategy, an algorithm that fleshes out the details of the strategy for subgames that are reached during play, and a self-improver algorithm that fixes potential weaknesses that opponents have identified in the blueprint strategy.
2018-02	**Adversarial Examples that Fool both Computer Vision and Time-Limited Humans** Gamaleldin F. Elsayed, Shreya Shankar, Brian Cheung et al. Latest version is dated 2018-05-22 https://arxiv.org/abs/1802.08195	Machine learning models are vulnerable to adversarial examples: small changes to images can cause computer vision models to make mistakes such as identifying a school bus as an ostrich. However, it is still an open question whether humans are prone to similar mistakes. Here, we address this question by leveraging recent techniques that transfer adversarial examples from computer vision models with known parameters and architecture to other models with unknown parameters and architecture, and by matching the initial processing of the human visual system. We find that adversarial examples that strongly transfer across computer vision models influence the classifications made by time-limited human observers.
2018-05 🐿	**To Build Truly Intelligent Machines, Teach Them Cause and Effect** Interview by Quanta Magazine with Kevin Hartnett https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/	"Judea Pearl, a pioneering figure in artificial intelligence, argues that AI has been stuck in a decades-long rut. His prescription for progress? Teach machines to understand the question why."

2018-06	**Relational inductive biases, deep learning, and graph networks** Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst et al. https://arxiv.org/abs/1806.01261	Artificial intelligence (AI) has undergone a renaissance recently, making major progress in key domains such as vision, language, control, and decision-making. This has been due, in part, to cheap data and cheap compute resources, which have fit the natural strengths of deep learning. However, many defining characteristics of human intelligence, which developed under much different pressures, remain out of reach for current approaches. In particular, generalizing beyond one's experiences--a hallmark of human intelligence from infancy--remains a formidable challenge for modern AI. The following is part position paper, part review, and part unification. We argue that combinatorial generalization must be a top priority for AI to achieve human-like abilities, and that structured representations and computations are key to realizing this objective. Just as biology uses nature and nurture cooperatively, we reject the false choice between "hand-engineering" and "end-to-end" learning, and instead advocate for an approach which benefits from their complementary strengths. We explore how using relational inductive biases within deep learning architectures can facilitate learning about entities, relations, and rules for composing them. We present a new building block for the AI toolkit with a strong relational inductive bias--the graph network--which generalizes and extends various approaches for neural networks that operate on graphs, and provides a straightforward interface for manipulating structured knowledge and producing structured behaviors. We discuss how graph networks can support relational reasoning and combinatorial generalization, laying the foundation for more sophisticated, interpretable, and flexible patterns of reasoning. As a companion to this paper, we have released an open-source software library for building graph networks, with demonstrations of how to use them in practice.
2018-09 🐿️	**Machine Learning Confronts the Elephant in the Room** Popular Science explainer at Quanta Magazine by Kevin Hartnett https://www.quantamagazine.org/machine-learning-confronts-the-elephant-in-the-room-20180920/	"A visual prank exposes an Achilles' heel of computer vision systems: Unlike humans, they can't do a double take."
2018-10	**BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding** Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova https://arxiv.org/abs/1810.04805 Latest version: 2019-05-24	🦉 **Introduces BERT, a breakthrough in natural language processing (NLP) that significantly improved the performance of language models across a wide range of NLP tasks.** We introduce a new language representation model called BERT, which stands for Bidirectional Encoder Representations from Transformers. Unlike recent language representation models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.
2018-10 📘	**Deep Reinforcement Learning** Yuxi Li https://arxiv.org/abs/1810.06339 ***150 pages of goodness!!***	We discuss deep reinforcement learning in an overview style. We draw a big picture, filled with details. We discuss six core elements, six important mechanisms, and twelve applications, focusing on contemporary work, and in historical contexts. We start with background of artificial intelligence, machine learning, deep learning, and reinforcement learning (RL), with resources. Next we discuss RL core elements, including value function, policy, reward, model, exploration vs. exploitation, and representation. Then we discuss important mechanisms for RL, including attention and memory, unsupervised learning, hierarchical RL, multi-agent RL, relational RL, and learning to learn. After that, we discuss RL applications, including games, robotics, natural language processing (NLP), computer vision, finance, business management, healthcare, education, energy, transportation, computer systems, and, science, engineering, and art. Finally we summarize briefly, discuss challenges and opportunities, and close with an epilogue.
2018-11 📙	**Reinforcement Learning, second edition: An Introduction** (Adaptive Computation and Machine Learning series) Richard S. Sutton, Andrew G. Barto MIT Press https://mitpressbookstore.mit.edu/book/9780262039246	Reinforcement learning, one of the most active research areas in artificial intelligence, is a computational approach to learning whereby an agent tries to maximize the total amount of reward it receives while interacting with a complex, uncertain environment. In *Reinforcement Learning*, Richard Sutton and Andrew Barto provide a clear and simple account of the field's key ideas and algorithms. This second edition has been significantly expanded and updated, presenting new topics and updating coverage of other topics. Like the first edition, this second edition focuses on core online learning algorithms, with the more mathematical material set off in shaded boxes. Part I covers as much of reinforcement learning as possible without going beyond the tabular case for which exact solutions can be found. Many algorithms presented in this part are new to the second edition, including UCB, Expected Sarsa, and Double Learning. Part II extends these ideas to function approximation, with new sections on such topics as artificial neural networks and the Fourier basis, and offers expanded treatment of off-policy learning and policy-gradient methods. Part III has new chapters on reinforcement learning's relationships to psychology and neuroscience, as well as an updated case-studies chapter including AlphaGo and AlphaGo Zero, Atari game playing, and IBM Watson's wagering strategy. The final chapter discusses the future societal impacts of reinforcement learning.
2018-11	**The Future of Artificial Intelligence in China** Jun Zhu, Tiejun Huang, Wenguang Chen, Wen Gao Appears in: "Communications of the ACM", November 2018, China Region Special Section: Hot Topics https://cacm.acm.org/research/the-future-of-artificial-intelligence-in-china/ https://dl.acm.org/doi/pdf/10.1145/3239540	
2019		
2019-01	**Where We See Shapes, AI Sees Textures** Popular Science explainer at Quanta Magazine by Jordana Cepelewicz https://www.quantamagazine.org/where-we-see-shapes-ai-sees-textures-20190701/	"To researchers' surprise, deep learning vision algorithms often fail at classifying images because they mostly take cues from textures, not shapes."
2019-01 📘	**A Comprehensive Survey on Graph Neural Networks** Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, Philip S. Yu All Aut	Deep learning has revolutionized many machine learning tasks in recent years, ranging from image classification and video processing to speech recognition and natural language understanding. The data in these tasks are typically represented in the Euclidean space. However, there is an increasing number of applications, where data are generated from non-Euclidean domains and are represented as graphs with complex relationships and interdependency between objects. The complexity of

	Published in: IEEE Transactions on Neural Networks and Learning Systems (Volume: 32, Issue: 1, January 2021) https://ieeexplore.ieee.org/document/9046288 https://arxiv.org/abs/1901.00596 Latest version: 2019-12-14	graph data has imposed significant challenges on the existing machine learning algorithms. Recently, many studies on extending deep learning approaches for graph data have emerged. In this article, we provide a comprehensive overview of graph neural networks (GNNs) in data mining and machine learning fields. We propose a new taxonomy to divide the state-of-the-art GNNs into four categories, namely, recurrent GNNs, convolutional GNNs, graph autoencoders, and spatial–temporal GNNs. We further discuss the applications of GNNs across various domains and summarize the open-source codes, benchmark data sets, and model evaluation of GNNs. Finally, we propose potential research directions in this rapidly growing field.
2019-01	**Foundations Built for a General Theory of Neural Networks** Popular Science explainer at Quanta Magazine by Kevin Hartnett https://www.quantamagazine.org/foundations-built-for-a-general-theory-of-neural-networks-20190131/	"Neural networks can be as unpredictable as they are powerful. Now mathematicians are beginning to reveal how a neural network's form will influence its function."
2019-02	**Language Models are Unsupervised Multitask Learners** Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf	🦉 **"The GPT-2 paper": A description of GPT-2 as used in the "Talk to Transformer" application.** Natural language processing tasks, such as question answering, machine translation, reading comprehension, and summarization, are typically approached with supervised learning on task-specific datasets. We demonstrate that language models begin to learn these tasks without any explicit supervision when trained on a new dataset of millions of webpages called WebText. When conditioned on a document plus questions, the answers generated by the language model reach 55 F1 on the CoQA dataset - matching or exceeding the performance of 3 out of 4 baseline systems without using the 127,000+ training examples. The capacity of the language model is essential to the success of zero-shot task transfer and increasing it improves performance in a log-linear fashion across tasks. Our largest model, GPT-2, is a 1.5B parameter Transformer that achieves state of the art results on 7 out of 8 tested language modeling datasets in a zero-shot setting but still underfits WebText. Samples from the model reflect these improvements and contain coherent paragraphs of text. These findings suggest a promising path towards building language processing systems which learn to perform tasks from their naturally occurring demonstrations
2019-03	**The Seven Tools of Causal Inference, with Reflections on Machine Learning** Judea Pearl Appears: in "Communications of the ACM", March 2019 https://cacm.acm.org/research/the-seven-tools-of-causal-inference-with-reflections-on-machine-learning/	🦉 **This is a modified discussion of "Theoretical Impediments to Machine Learning" of 2018-01** The kind of causal inference seen in natural human thought can be “algorithmitized” to help produce human-level machine intelligence
2019-03 📅	**Fathers of the Deep Learning Revolution Receive ACM A.M. Turing Award 2018** https://www.acm.org/media-center/2019/march/turing-award-2018	Bengio, Hinton and LeCun Ushered in Major Breakthroughs in Artificial Intelligence
2019-06	**Reaching New Heights with Artificial Neural Networks** https://cacm.acm.org/opinion/reaching-new-heights-with-artificial-neural-networks/?utm_source=chatgpt.com Appears in "Communications of the ACM", June 2019 https://cacm.acm.org/issue/june-2019/	Turing Award 2018 discussion "ACM A.M. Turing Award recipients Yoshua Bengio, Geoffrey Hinton, and Yann LeCun on the promise of neural networks, the need for new paradigms, and the concept of making technology accessible to all.
2019-06	**Research for Practice: Troubling Trends in Machine-Learning Scholarship** Appears in "Communications of the ACM", June 2019 https://cacm.acm.org/practice/research-for-practice-12/	"Some ML papers suffer from flaws that could mislead the public and stymie future research."
2019-06	**The Challenge of Crafting Intelligible Intelligence** By Daniel S. Weld and Gagan Bansal Appears in "Communications of the ACM", June 2019 https://cacm.acm.org/research/the-challenge-of-crafting-intelligible-intelligence/	"To trust the behavior of complex AI algorithms, especially in mission-critical settings, they must be made intelligible."
2019-06	**Adaptive activation functions accelerate convergence in deep and physics-informed neural networks** Ameya D. Jagtapa, George Em Karniadakisa https://arxiv.org/abs/1906.01170	We employ adaptive activation functions for regression in deep and physics-informed neural networks (PINNs) to approximate smooth and discontinuous functions as well as solutions of linear and nonlinear partial differential equations. In particular, we solve the nonlinear Klein-Gordon equation, which has smooth solutions, the nonlinear Burgers equation, which can admit high gradient solutions, and the Helmholtz equation. We introduce a scalable hyperparameter in the activation function, which can be optimized to achieve best performance of the network as it changes dynamically the topology of the loss function involved in the optimization process. The adaptive activation function has better learning capabilities than the traditional one (fixed activation) as it improves greatly the convergence rate, especially at early training, as well as the solution accuracy. To better understand the learning process, we plot the neural network solution in the frequency domain to examine how the network captures successively different frequency bands present in the solution. We consider both forward problems, where the approximate solutions are obtained, as well as inverse problems, where parameters involved in the governing equation are identified. Our simulation results show that the proposed method is a very simple and effective approach to increase the efficiency, robustness and accuracy of the neural network approximation of nonlinear functions as well as solutions of partial differential equations, especially for forward problems.
2019-07	**Machines Beat Humans on a Reading Test. But Do They Understand?** Popular Science explainer at Quanta Magazine by John Pavlus https://www.quantamagazine.org/machines-beat-humans-on-a-reading-test-but-do-	"A tool known as BERT can now beat humans on advanced reading-comprehension tests. But it's also revealed how far AI has to go."

	they-understand-20191017/	
2019-10	**Deep Learning: Our Miraculous Year 1990-1991** Jürgen Schmidhuber The HTML page was inaugurated on 2019-10-04 and is updated from time to time: https://people.idsia.ch/~juergen/deep-learning-miraculous-year-1990-1991.html The latest PDF version at the ArxiV is dated 2022-12-28: https://arxiv.org/abs/2005.05744 Also see: **My First Deep Learning System of 1991 + Deep Learning Timeline 1960-2013** https://people.idsia.ch/~juergen/firstdeeplearner.html	🦉 **A timeline of "Deep Learning" history as seen from TU München and IDSIA Switzerland.** "In 2020-2021, we celebrated that many of the basic ideas behind the deep learning revolution were published three decades ago within fewer than 12 months in our "Annus Mirabilis" or "Miraculous Year" 1990-1991 at TU Munich. Back then, few people were interested, but a quarter century later, neural networks based on these ideas were on over 3 billion devices such as smartphones, and used many billions of times per day, consuming a significant fraction of the world's compute."
2019-11	**Intelligence May Not Be Computable** Peter J. Denning, Ted G. Lewis http://denninginstitute.com/pjd/PUBS/amsci-2019-ai-hierachy.pdf Appears in: American Scientist November/December 2019	A hierarchy of artificial intelligence machines ranked by their learning power shows their abilities—and their limits. See also: **Interactive Cognitive Systems and Social Intelligence, 2017-07**
2019-12	**Characterizing the Decision Boundary of Deep Neural Networks** Hamid Karimi, Tyler Derr, Jiliang Tang https://arxiv.org/abs/1912.11460 Latest version: 2020-06-03	Deep neural networks and in particular, deep neural classifiers have become an integral part of many modern applications. Despite their practical success, we still have limited knowledge of how they work and the demand for such an understanding is evergrowing. In this regard, one crucial aspect of deep neural network classifiers that can help us deepen our knowledge about their decision-making behavior is to investigate their decision boundaries. Nevertheless, this is contingent upon having access to samples populating the areas near the decision boundary. To achieve this, we propose a novel approach we call Deep Decision boundary Instance Generation (DeepDIG). DeepDIG utilizes a method based on adversarial example generation as an effective way of generating samples near the decision boundary of any deep neural network model. Then, we introduce a set of important principled characteristics that take advantage of the generated instances near the decision boundary to provide multifaceted understandings of deep neural networks. We have performed extensive experiments on multiple representative datasets across various deep neural network models and characterized their decision boundaries. The code is publicly available at this https URL.
2020		
2020-01	**The unreasonable effectiveness of deep learning in artificial intelligence** Terrence J. Sejnowski https://www.pnas.org/doi/10.1073/pnas.1907373117	A philosophical article on why this stuff is working so well
2020-01 🐿	**Artificial Intelligence Will Do What We Ask. That's a Problem.** Popular Science explainer at Quanta Magazine by Natalie Wolchover https://www.quantamagazine.org/artificial-intelligence-will-do-what-we-ask-thats-a-problem-20200130/	"By teaching machines to understand our true desires, one scientist hopes to avoid the potentially disastrous consequences of having them do what we command."
2020-01	**Techniques for Interpretable Machine Learning** Mengnan Du, Ninghao Liu, and Xia Hu Appears in: "Communications of the ACM", January 2020 https://cacm.acm.org/research/techniques-for-interpretable-machine-learning/	Uncovering the mysterious ways machine learning models make decisions.
2020-02	**Emergent Tool Use from Multi-Agent Autocurricula** OpenAI / Google Brain https://arxiv.org/abs/1909.07528	Through multi-agent competition, the simple objective of hide-and-seek, and standard reinforcement learning algorithms at scale, we find that agents create a self-supervised autocurriculum inducing multiple distinct rounds of emergent strategy, many of which require sophisticated tool use and coordination. We find clear evidence of six emergent phases in agent strategy in our environment, each of which creates a new pressure for the opposing team to adapt; for instance, agents learn to build multi-object shelters using moveable boxes which in turn leads to agents discovering that they can overcome obstacles using ramps. We further provide evidence that multi-agent competition may scale better with increasing environment complexity and leads to behavior that centers around far more human-relevant skills than other self-supervised reinforcement learning methods such as intrinsic motivation. Finally, we propose transfer and fine-tuning as a way to quantitatively evaluate targeted capabilities, and we compare hide-and-seek agents to both intrinsic motivation and random initialization baselines in a suite of domain-specific intelligence tests.
2020-04 🐿	**Common Sense Comes Closer to Computers** Popular Science explainer at Quanta Magazine by John Pavlus https://www.quantamagazine.org/common-sense-comes-closer-to-computers-20200430/	"The problem of common-sense reasoning has plagued the field of artificial intelligence for over 50 years. Now a new approach, borrowing from two disparate lines of thinking, has made important progress."
2020-05	**Language Models are Few-Shot Learners** Tom B. Brown, Benjamin Mann, Nick Ryder et al. https://arxiv.org/abs/2005.14165	🦉 **Introduction of GPT-3, a powerful autoregressive language model with 175 billion parameters. This paper:** • **shifted the paradigm in NLP from task-specific models to general-purpose language models.** • **sparked the era of foundation models, leading to further advancements like ChatGPT, GPT-4, and other large-scale AI models**. • **provided a proof of concept that increasing scale leads to emergent capabilities in AI.**

		Recent work has demonstrated substantial gains on many NLP tasks and benchmarks by pre-training on a large corpus of text followed by fine-tuning on a specific task. While typically task-agnostic in architecture, this method still requires task-specific fine-tuning datasets of thousands or tens of thousands of examples. By contrast, humans can generally perform a new language task from only a few examples or from simple instructions - something which current NLP systems still largely struggle to do. Here we show that scaling up language models greatly improves task-agnostic, few-shot performance, sometimes even reaching competitiveness with prior state-of-the-art fine-tuning approaches.
2020-06	**Design Lessons From AI's Two Grand Goals: Human Emulation and Useful Applications** Ben Shneiderman Appears in: "IEEE Transactions on Technology and Society, Vol. 1, No. 2, June 2020" https://ieeexplore.ieee.org/document/9088114 (open access)	Researchers' goals shape the questions they raise, collaborators they choose, methods they use, and outcomes of their work. This article offers a fresh vision of artificial intelligence (AI) research by suggesting a simplification to two goals: 1. emulation to understand human abilities to build systems that perform tasks as well as or better than humans; and 2. application of AI methods to build widely used products and services. Researchers and developers for each goal can fruitfully work along their desired paths, but this article is intended to limit the problems that arise when assumptions from one goal are used to drive work on the other goal. For example, autonomous humanoid robots are prominent with emulation researchers, but application developers avoid them, in favor of tool-like appliances or teleoperated devices for widely used commercial products and services. This article covers four such mismatches in goals that affect AI-guided application development: 1. intelligent agent or powerful tool; 2. simulated teammate or teleoperated device; 3. autonomous system or supervisory control; and 4. humanoid robot or mechanoid appliance. This article clarifies these mismatches to facilitate the discovery of workable compromise designs that will accelerate human-centered AI applications research. A greater emphasis on human-centered AI could reduce AI's existential threats and increase benefits for users and society, such as in business, education, healthcare, environmental preservation, and community safety.
2020-09	**Analysis of Generalizability of Deep Neural Networks Based on the Complexity of Decision Boundary** Shuyue Guan, Murray Loew https://arxiv.org/abs/2009.07974	For supervised learning models, the analysis of generalization ability (generalizability) is vital because the generalizability expresses how well a model will perform on unseen data. Traditional generalization methods, such as the VC dimension, do not apply to deep neural network (DNN) models. Thus, new theories to explain the generalizability of DNNs are required. In this study, we hypothesize that the DNN with a simpler decision boundary has better generalizability by the law of parsimony (Occam's Razor). We create the decision boundary complexity (DBC) score to define and measure the complexity of decision boundary of DNNs. The idea of the DBC score is to generate data points (called adversarial examples) on or near the decision boundary. Our new approach then measures the complexity of the boundary using the entropy of eigenvalues of these data. The method works equally well for high-dimensional data. We use training data and the trained model to compute the DBC score. And, the ground truth for model's generalizability is its test accuracy. Experiments based on the DBC score have verified our hypothesis. The DBC is shown to provide an effective method to measure the complexity of a decision boundary and gives a quantitative measure of the generalizability of DNNs.
2020-11	**The Complexity of Gradient Descent: CLS = PPAD ∩ PLS** John Fearnley, Paul W. Goldberg, Alexandros Hollender, Rahul Savani https://arxiv.org/abs/2011.01929 Latest version: 2rd March, 2023	We study search problems that can be solved by performing Gradient Descent on a bounded convex polytopal domain and show that this class is equal to the intersection of two well-known classes: PPAD and PLS. As our main underlying technical contribution, we show that computing a Karush-Kuhn-Tucker (KKT) point of a continuously differentiable function over the domain [0, 1]2 is PPAD ∩ PLS-complete. This is the first natural problem to be shown complete for this class. Our results also imply that the class CLS (Continuous Local Search) – which was defined by Daskalakis and Papadimitriou as a more "natural" counterpart to PPAD ∩ PLS and contains many interesting problems – is itself equal to PPAD ∩ PLS.
2020-12	**Neurosymbolic AI: The 3rd Wave** Arthur a'Avila Garcez, Luís C. Lamb https://arxiv.org/abs/2012.05876	Current advances in Artificial Intelligence (AI) and Machine Learning (ML) have achieved unprecedented impact across research communities and industry. Nevertheless, concerns about trust, safety, interpretability and accountability of AI were raised by influential thinkers. Many have identified the need for well-founded knowledge representation and reasoning to be integrated with deep learning and for sound explainability. Neural-symbolic computing has been an active area of research for many years seeking to bring together robust learning in neural networks with reasoning and explainability via symbolic representations for network models. In this paper, we relate recent and early research results in neurosymbolic AI with the objective of identifying the key ingredients of the next wave of AI systems. We focus on research that integrates in a principled way neural network-based learning with symbolic knowledge representation and logical reasoning. The insights provided by 20 years of neural-symbolic computing are shown to shed new light onto the increasingly prominent role of trust, safety, interpretability and accountability of AI. We also identify promising directions and challenges for the next decade of AI research from the perspective of neural-symbolic systems.
2021		
2021 📙	**Linguistics for the Age of AI** Marjorie McShane, Sergei Nirenburg Open-access book by MIT Press https://direct.mit.edu/books/oa-monograph/5042/Linguistics-for-the-Age-of-AI	*Description of the book:* One of the original goals of artificial intelligence research was to endow intelligent agents with human-level natural language capabilities. Recent AI research, however, has focused on applying statistical and machine learning approaches to big data rather than attempting to model what people do and how they do it. In this book, Marjorie McShane and Sergei Nirenburg return to the original goal of recreating human-level intelligence in a machine. They present a human-inspired, linguistically sophisticated model of language understanding for intelligent agent systems that emphasizes meaning—the deep, context-sensitive meaning that a person derives from spoken or written language. With Linguistics for the Age of AI, McShane and Nirenburg offer a roadmap for creating language-endowed intelligent agents (LEIAs) that can understand,explain, and learn. They describe the language-understanding capabilities of LEIAs from the perspectives of cognitive modeling and system building, emphasizing "actionability"—which involves achieving interpretations that are sufficiently deep, precise, and confident to support reasoning about action. After detailing their microtheories for topics such as semantic analysis, basic coreference, and situational reasoning, McShane and Nirenburg turn to agent applications developed using those microtheories and evaluations of a LEIA's language understanding capabilities. McShane and Nirenburg argue that the only way to achieve human-level language understanding by machines is to place linguistics front and center, using statistics

		and big data as contributing resources. They lay out a long-term research program that addresses linguistics and real-world reasoning together, within a comprehensive cognitive architecture. The open access edition of this book was made possible by generous funding from Arcadia – a charitable fund of Lisbet Rausing and Peter Baldwin.
2021-01 📘	**A Survey of Complex-Valued Neural Networks** Joshua Bassey, Lijun Qian, Xianfang Li https://arxiv.org/abs/2101.12249	Artificial neural networks (ANNs) based machine learning models and especially deep learning models have been widely applied in computer vision, signal processing, wireless communications, and many other domains, where complex numbers occur either naturally or by design. However, most of the current implementations of ANNs and machine learning frameworks are using real numbers rather than complex numbers. There are growing interests in building ANNs using complex numbers, and exploring the potential advantages of the so-called complex-valued neural networks (CVNNs) over their real-valued counterparts. In this paper, we discuss the recent development of CVNNs by performing a survey of the works on CVNNs in the literature. Specifically, a detailed review of various CVNNs in terms of activation function, learning and optimization, input and output representations, and their applications in tasks such as signal processing and computer vision are provided, followed by a discussion on some pertinent challenges and future research directions.
2021-01 🐿️	**Same or Different? The Question Flummoxes Neural Networks.** Popular Science explainer at Quanta Magazine by John Pavlus https://www.quantamagazine.org/same-or-different-ai-cant-tell-20210623/	"For all their triumphs, AI systems can't seem to generalize the concepts of 'same' and 'different.' Without that, researchers worry, the quest to create truly intelligent machines may be hopeless."
2021-04	**Why AI is Harder Than We Think** Melanie Mitchell https://arxiv.org/abs/2104.12871	Since its beginning in the 1950s, the field of artificial intelligence has cycled several times between periods of optimistic predictions and massive investment ("AI spring") and periods of disappointment, loss of confidence, and reduced funding ("AI winter"). Even with today's seemingly fast pace of AI breakthroughs, the development of long-promised technologies such as self-driving cars, housekeeping robots, and conversational companions has turned out to be much harder than many people expected. One reason for these repeating cycles is our limited understanding of the nature and complexity of intelligence itself. In this paper I describe four fallacies in common assumptions made by AI researchers, which can lead to overconfident predictions about the field. I conclude by discussing the open questions spurred by these fallacies, including the age-old challenge of imbuing machines with humanlike common sense.
2021-04	**The Best of NLP** Natural language processing delves more deeply into its knowledge gap (review) Chris Edwards Appears in: "Communications of the ACM", April 2021 https://dl.acm.org/doi/10.1145/3449049	
2021-04	**Transformers Aftermath: Current Research and Rising Trends** Eduardo Souza Dos Reis, Cristiano André Da Costa, Diórgenes Eugênio Da Silveira et al. Appears in: "Communications of the ACM", April 2021 https://dl.acm.org/doi/10.1145/3430937	"Attention, particularly self-attention, is a standard in current NLP literature, but to achieve meaningful models, attention is not enough."
2021-05 🐿️	**Douglas Hofstadter's Eight Abilities for Intelligence: Humans vs. Artificial Intelligence** Paul Austin Murphy https://becominghuman.ai/douglas-hofstadters-eight-abilities-for-intelligence-humans-vs-artificial-intelligence-b6953af1ace4	Discusses the following capabilities we would like to see: 1. to respond to situations very flexibly; 2. to take advantage of fortuitous circumstances; 3. to make sense out of ambiguous or contradictory messages; 4. to recognise the relative importance of different elements of a situation; 5. to find similarities between situations despite differences which may separate them; 6. to draw distinctions between situations despite similarities which may link them; 7. to synthesize new concepts by taking old concepts and putting them together in new ways; 8. to come up with ideas which are novel.
2021-06	**LoRA: Low-Rank Adaptation of Large Language Models** Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen https://arxiv.org/abs/2106.09685	🦉 **Presents a novel approach for efficiently adapting large pre-trained models without the need for extensive fine-tuning.** An important paradigm of natural language processing consists of large-scale pre-training on general domain data and adaptation to particular tasks or domains. As we pre-train larger models, full fine-tuning, which retrains all model parameters, becomes less feasible. Using GPT-3 175B as an example -- deploying independent instances of fine-tuned models, each with 175B parameters, is prohibitively expensive. We propose Low-Rank Adaptation, or LoRA, which freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture, greatly reducing the number of trainable parameters for downstream tasks. Compared to GPT-3 175B fine-tuned with Adam, LoRA can reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times. LoRA performs on-par or better than fine-tuning in model quality on RoBERTa, DeBERTa, GPT-2, and GPT-3, despite having fewer trainable parameters, a higher training throughput, and, unlike adapters, no additional inference latency. We also provide an empirical investigation into rank-deficiency in language model adaptation, which sheds light on the efficacy of LoRA. We release a package that facilitates the integration of LoRA with PyTorch models and provide our implementations and model checkpoints for RoBERTa, DeBERTa, and GPT-2 at this https
2021-07	**Deep Learning for AI: Turing lecture for the 2018 ACM A.M. Turing Award** Yoshua Bengio, Yann LeCun, and Geoffrey Hinton https://cacm.acm.org/research/deep-learning-for-ai/	"How can neural networks learn the rich internal representations required for difficult tasks such as recognizing objects or understanding language?"

2021-07	**The Computer Scientist Training AI to Think With Analogies** Popular Science explainer at Quanta Magazine by John Pavlus https://www.quantamagazine.org/melanie-mitchell-trains-ai-to-think-with-analogies-20210714/	"Melanie Mitchell has worked on digital minds for decades. She says they'll never truly be like ours until they can make analogies."
2021-09	**How Computationally Complex Is a Single Neuron?** Popular Science explainer at Quanta Magazine by Allison Whitten https://www.quantamagazine.org/how-computationally-complex-is-a-single-neuron-20210902/	"Computational neuroscientists taught an artificial neural network to imitate a biological neuron. The result offers a new way to think about the complexity of single brain cells." The answer seems to be about 1000 artificial neurons
2021-10	**The Uselessness of Useful Knowledge / Is artificial intelligence the new alchemy?** Popular Science explainer at Quanta Magazine by Robbert Dijkgraaf https://www.quantamagazine.org/science-has-entered-a-new-era-of-alchemy-good-20211020/	"Today's powerful but little-understood artificial intelligence breakthroughs echo past examples of unexpected scientific progress."
2021-10	**A New Link to an Old Model Could Crack the Mystery of Deep Learning** Popular Science explainer at Quanta Magazine by Anil Ananthaswamy https://www.quantamagazine.org/a-new-link-to-an-old-model-could-crack-the-mystery-of-deep-learning-20211011/	"To help them explain the shocking success of deep neural networks, researchers are turning to older but better-understood models of machine learning."
2021-10	**Trustworthy AI** Jeannette M. Wing Appears in: "Communications of the ACM", October 2021 https://dl.acm.org/doi/abs/10.1145/3448248	The pursuit of responsible AI raises the ante on both the trustworthy computing and formal methods communities
2021-10	**Learning High-Speed Flight in the Wild** Antonia Loquerico, Elia Kuafmann, René Ranftl, Matthias Müller, Vladen Koltun, Davide Scaramuzza Appears in: "Science Robotics" Vol. 6, No. 59 (October 6, 2021) https://www.science.org/doi/10.1126/scirobotics.abg5810 Alternatively: https://arxiv.org/abs/2110.05113 *See also:* **Autonomous Racing Drones Dodge Through Forests at 40 kph** "Training in simulation gives these drones impressive flying skills" Appears in: "IEEE Spectrum", 2021-10-07 https://spectrum.ieee.org/racing-drone **Superhuman Speed: How Autonomous Drones Beat the Best Human Racers** "Processing on the fly gives extreme drones the edge" Appears in: "IEEE Spectrum", 2023-08-31 https://spectrum.ieee.org/ai-drone-racing	Quadrotors are agile. Unlike most other machines, they can traverse extremely complex environments at high speeds. To date, only expert human pilots have been able to fully exploit their capabilities. Autonomous operation with onboard sensing and computation has been limited to low speeds. State-of-the-art methods generally separate the navigation problem into subtasks: sensing, mapping, and planning. Although this approach has proven successful at low speeds, the separation it builds upon can be problematic for high-speed navigation in cluttered environments. The subtasks are executed sequentially, leading to increased processing latency and a compounding of errors through the pipeline. Here we propose an end-to-end approach that can autonomously fly quadrotors through complex natural and human-made environments at high speeds, with purely onboard sensing and computation. The key principle is to directly map noisy sensory observations to collision-free trajectories in a receding-horizon fashion. This direct mapping drastically reduces processing latency and increases robustness to noisy and incomplete perception. The sensorimotor mapping is performed by a convolutional network that is trained exclusively in simulation via privileged learning: imitating an expert with access to privileged information. By simulating realistic sensor noise, our approach achieves zero-shot transfer from simulation to challenging real-world environments that were never experienced during training: dense forests, snow-covered terrain, derailed trains, and collapsed buildings. Our work demonstrates that end-to-end policies trained in simulation enable high-speed autonomous flight through challenging environments, outperforming traditional obstacle avoidance pipelines.
2021-12	**What Does It Mean for AI to Understand?** Popular Science explainer at Quanta Magazine by Melanie Mitchell https://www.quantamagazine.org/what-does-it-mean-for-ai-to-understand-20211216/	"It's simple enough for AI to seem to comprehend data, but devising a true test of a machine's knowledge has proved difficult."
2021-12	**Datasheet for Datasets** "Documentation to facilitate communication between dataset creators and consumers." Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, Kate Crawford Appears in: "Communications of the ACM", December 2021 https://arxiv.org/abs/1803.09010 https://cacm.acm.org/research/datasheets-for-datasets/	Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of static datasets. The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases. Mismatches like this can have especially severe consequences when machine learning models are used in high-stakes domains, such as criminal justice, hiring critical infrastructure, and finance. Even in other domains, mismatches may lead to loss of revenue or public relations setbacks. Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in training datasets. For these and other reasons, the World Economic Forum suggests all entities should document the provenance, creation, and use of machine learning datasets to avoid discriminatory outcomes
2022		

2022-01	**Researchers Build AI That Builds AI** Popular Science explainer at Quanta Magazine by Anil Ananthaswamy https://www.quantamagazine.org/researchers-build-ai-that-builds-ai-20220125/	"By using hypernetworks, researchers can now preemptively fine-tune artificial neural networks, saving some of the time and expense of training."
2022-01	**Creating a 30-Million-Rule System: MCC and Cycorp** Douglas Lenat https://ieeexplore.ieee.org/document/9713910 Published in: IEEE Annals of the History of Computing (Volume: 44, Issue: 1, 01 Jan.-March 2022)	🦉 **Cyc Review** Hard-won discoveries led to early expert systems (ESs) successes, then overhyping, and then disillusionment. The bottleneck was infrastructure: limited expressivity representation languages, inefficient inference engines, inadequate ontologies, as well as lack of common-sense, general theories of the world, and argumentation and context mechanisms. At Microelectronics and Computer Technology Corporation and then at Cycorp, we have systematically codified much of the “obvious” knowledge of the world that one rarely articulates since “everyone” of course already knows it. Lack of that solid infrastructure limits AIs' trustworthiness: they make mistakes no human would make, and they cannot explain their reasoning. This is the story of my ESs experience and how 50 years of lessons learned led my team to steadily and successfully construct an enormous knowledge-based system that avoids such brittleness.
2022-03	**Will Transformers Take Over Artificial Intelligence?** Popular Science explainer at Quanta Magazine by Stephen Ornes https://www.quantamagazine.org/will-transformers-take-over-artificial-intelligence-20220310/	"A simple algorithm that revolutionized how neural networks approach language is now taking on vision as well. It may not stop there."
2022-03	**Training language models to follow instructions with human feedback** *"The Instruct-GPT paper"* https://arxiv.org/abs/2203.02155	Making language models bigger does not inherently make them better at following a user's intent. For example, large language models can generate outputs that are untruthful, toxic, or simply not helpful to the user. In other words, these models are not aligned with their users. In this paper, we show an avenue for aligning language models with user intent on a wide range of tasks by fine-tuning with human feedback. Starting with a set of labeler-written prompts and prompts submitted through the OpenAI API, we collect a dataset of labeler demonstrations of the desired model behavior, which we use to fine-tune GPT-3 using supervised learning. We then collect a dataset of rankings of model outputs, which we use to further fine-tune this supervised model using reinforcement learning from human feedback. We call the resulting models InstructGPT. In human evaluations on our prompt distribution, outputs from the 1.3B parameter InstructGPT model are preferred to outputs from the 175B GPT-3, despite having 100x fewer parameters. Moreover, InstructGPT models show improvements in truthfulness and reductions in toxic output generation while having minimal performance regressions on public NLP datasets. Even though InstructGPT still makes simple mistakes, our results show that fine-tuning with human feedback is a promising direction for aligning language models with human intent.
2022-05	**Information Theory as a Bridge Between Language Function and Language Form** Richard Futrell and Michael Hahn https://www.frontiersin.org/journals/communication/articles/10.3389/fcomm.2022.657725/full	Formal and functional theories of language seem disparate, because formal theories answer the question of what a language is, while functional theories answer the question of what functions it serves. We argue that information theory provides a bridge between these two approaches, via a principle of minimization of complexity under constraints. Synthesizing recent work, we show how information-theoretic characterizations of functional complexity lead directly to mathematical descriptions of the forms of possible languages, in terms of solutions to constrained optimization problems. We show how certain linguistic descriptive formalisms can be recovered as solutions to such problems. Furthermore, we argue that information theory lets us define complexity in a way which has minimal dependence on the choice of theory or descriptive formalism. We illustrate this principle using recently-obtained results on universals of word and morpheme order
2022-05	**A very preliminary analysis of DALL-E 2** Gary Marcus, Ernest Davis, Scott Aaronson https://arxiv.org/abs/2204.13807	The DALL-E 2 system generates original synthetic images corresponding to an input text as caption. We report here on the outcome of fourteen tests of this system designed to assess its common sense, reasoning and ability to understand complex texts. All of our prompts were intentionally much more challenging than the typical ones that have been showcased in recent weeks. Nevertheless, for 5 out of the 14 prompts, at least one of the ten images fully satisfied our requests. On the other hand, on no prompt did all of the ten images satisfy our requests.
2022-06	**Discovering the Hidden Vocabulary of DALL-E 2** Giannis Daras, Alexandros G. Dimakis https://arxiv.org/abs/2206.00169	We discover that DALLE-2 seems to have a hidden vocabulary that can be used to generate images with absurd prompts. For example, it seems that \texttt{Apoploe vesrreaitais} means birds and \texttt{Contarra ccetnxniams luryca tanniounons} (sometimes) means bugs or pests. We find that these prompts are often consistent in isolation but also sometimes in combinations. We present our black-box method to discover words that seem random but have some correspondence to visual concepts. This creates important security and interpretability challenges.
2022-06	**A physics-based digital twin for model predictive control of autonomous unmanned aerial vehicle landing** Andrew McClellan, Joseph Lorenzetti, Marco Pavone, Charbel Farhat https://royalsocietypublishing.org/doi/full/10.1098/rsta.2021.0204	This paper proposes a two-level, data-driven, digital twin concept for the autonomous landing of aircraft, under some assumptions. It features a digital twin instance (DTI) for model predictive control (MPC); and an innovative, real-time, digital twin prototype for fluid–structure interaction and flight dynamics to inform it. The latter digital twin is based on the linearization about a pre-designed glideslope trajectory of a high-fidelity, viscous, nonlinear computational model for flight dynamics; and its projection onto a low-dimensional approximation subspace to achieve real-time performance, while maintaining accuracy. Its main purpose is to predict in real time, during flight, the state of an aircraft and the aerodynamic forces and moments acting on it. Unlike static lookup tables or regression-based surrogate models based on steady-state wind tunnel data, the aforementioned real-time digital twin prototype allows the DTI for MPC to be informed by a truly dynamic flight model, rather than a less accurate set of steady-state aerodynamic force and moment data points. The paper describes in detail the construction of the proposed two-level digital twin concept and its verification by numerical simulation. It also reports on its preliminary flight validation in autonomous mode for an off-the-shelf unmanned aerial vehicle instrumented at Stanford University. This article is part of the theme issue ‘Data-driven prediction in dynamical systems’.
2022-07	**Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models** Numerous authors https://arxiv.org/abs/2206.04615 Latest version is date 2023-06-12	Language models demonstrate both quantitative improvement and new qualitative capabilities with increasing scale. Despite their potentially transformative impact, these new capabilities are as yet poorly characterized. In order to inform future research, prepare for disruptive new model capabilities, and ameliorate socially harmful effects, it is vital that we understand the present and near-future capabilities and limitations of language models. To address this challenge, we introduce the Beyond the Imitation Game benchmark (BIG-bench). BIG-bench currently consists of 204 tasks, contributed by 450 authors across 132 institutions. Task topics are diverse, drawing problems from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development, and beyond. BIG-bench focuses on tasks that are believed to be beyond the capabilities of current language models.

2022-09	**How Transformers Seem to Mimic Parts of the Brain** Popular Science explainer at Quanta Magazine by Stephen Ornes https://www.quantamagazine.org/how-ai-transformers-mimic-parts-of-the-brain-20220912/	"Neural networks originally designed for language processing turn out to be great models of how our brains understand places."
2022-09	**Neurosymbolic AI** Don Monroe Appears in: "Communications of the ACM", October 2022 https://dl.acm.org/doi/10.1145/3554918	"Combining neural networks with symbolic representations might make them more versatile and dependable."
2022-10	**Scaling Laws for Reward Model Overoptimization** Leo Gao, John Schulman, Jacob Hilton https://arxiv.org/abs/2210.10760	In reinforcement learning from human feedback, it is common to optimize against a reward model trained to predict human preferences. Because the reward model is an imperfect proxy, optimizing its value too much can hinder ground truth performance, in accordance with Goodhart's law. This effect has been frequently observed, but not carefully measured due to the expense of collecting human preference data. In this work, we use a synthetic setup in which a fixed "gold-standard" reward model plays the role of humans, providing labels used to train a proxy reward model. We study how the gold reward model score changes as we optimize against the proxy reward model using either reinforcement learning or best-of-n sampling. We find that this relationship follows a different functional form depending on the method of optimization, and that in both cases its coefficients scale smoothly with the number of reward model parameters. We also study the effect on this relationship of the size of the reward model dataset, the number of reward model and policy parameters, and the coefficient of the KL penalty added to the reward in the reinforcement learning setup. We explore the implications of these empirical results for theoretical considerations in AI alignment.
2022-12	**Talking About Large Language Models** Murray Shanahan https://arxiv.org/abs/2212.03551 (latest version 2023-02-16)	Thanks to rapid progress in artificial intelligence, we have entered an era when technology and philosophy intersect in interesting ways. Sitting squarely at the centre of this intersection are large language models (LLMs). The more adept LLMs become at mimicking human language, the more vulnerable we become to anthropomorphism, to seeing the systems in which they are embedded as more human-like than they really are. This trend is amplified by the natural tendency to use philosophically loaded terms, such as "knows", "believes", and "thinks", when describing these systems. To mitigate this trend, this paper advocates the practice of repeatedly stepping back to remind ourselves of how LLMs, and the systems of which they form a part, actually work. The hope is that increased scientific precision will encourage more philosophical nuance in the discourse around artificial intelligence, both within the field and in the public sphere.
2022-12	**Annotated History of Modern AI and Deep Learning** Juergen Schmidhuber https://people.idsia.ch/~juergen/deep-learning-history.html https://arxiv.org/abs/2212.11279	Machine learning is the science of credit assignment: finding patterns in observations that predict the consequences of actions and help to improve future performance. Credit assignment is also required for human understanding of how the world works, not only for individuals navigating daily life, but also for academic professionals like historians who interpret the present in light of past events. Here I focus on the history of modern artificial intelligence (AI) which is dominated by artificial neural networks (NNs) and deep learning, both conceptually closer to the old field of cybernetics than to what's been called AI since 1956 (e.g., expert systems and logic programming). A modern history of AI will emphasize breakthroughs outside of the focus of traditional AI text books, in particular, mathematical foundations of today's NNs such as the chain rule (1676), the first NNs (linear regression, circa 1800), and the first working deep learners (1965-). From the perspective of 2022, I provide a timeline of the -- in hindsight -- most important relevant events in the history of NNs, deep learning, AI, computer science, and mathematics in general, crediting those who laid foundations of the field. The text contains numerous hyperlinks to relevant overview sites from my AI Blog. It supplements my previous deep learning survey (2015) which provides hundreds of additional references. Finally, to round it off, I'll put things in a broader historic context spanning the time since the Big Bang until when the universe will be many times older than it is now.
2022-12	**BigText-QA: Question Answering over a Large-Scale Hybrid Knowledge Graph** Jingjing Xu, Maria Biryukov, Martin Theobald, Vinu Ellampallil Venugopal https://arxiv.org/abs/2212.05798	Answering complex questions over textual resources remains a challenge, particularly when dealing with nuanced relationships between multiple entities expressed within natural-language sentences. To this end, curated knowledge bases (KBs) like YAGO, DBpedia, Freebase, and Wikidata have been widely used and gained great acceptance for question-answering (QA) applications in the past decade. While these KBs offer a structured knowledge representation, they lack the contextual diversity found in natural-language sources. To address this limitation, BigText-QA introduces an integrated QA approach, which is able to answer questions based on a more redundant form of a knowledge graph (KG) that organizes both structured and unstructured (i.e., "hybrid") knowledge in a unified graphical representation. Thereby, BigText-QA is able to combine the best of both worlds–a canonical set of named entities, mapped to a structured background KB (such as YAGO or Wikidata), as well as an open set of textual clauses providing highly diversified relational paraphrases with rich context information. Our experimental results demonstrate that BigText-QA outperforms DrQA, a neural-network-based QA system, and achieves competitive results to QUEST, a graph-based unsupervised QA system.
2022-12	**Reasoning with Language Model Prompting: A Survey** Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, Huajun Chen https://arxiv.org/abs/2212.09597 Latest version: 2023-09-18	Reasoning, as an essential ability for complex problem-solving, can provide back-end support for various real-world applications, such as medical diagnosis, negotiation, etc. This paper provides a comprehensive survey of cutting-edge research on reasoning with language model prompting. We introduce research works with comparisons and summaries and provide systematic resources to help beginners. We also discuss the potential reasons for emerging such reasoning abilities and highlight future research directions. Resources are available at this https URL (updated periodically).
2023		
2023-01	**Neurosymbolic AI - Why, What, and How** Amit Sheth, Kaushik Roy, Manas Gaur https://arxiv.org/abs/2305.00813 Appears in "IEEE Intelligent Systems" May-June 2023 https://ieeexplore.ieee.org/document/10148662 (paywalled) https://www.computer.org/csdl/magazine/ex/2023/03/10148662/1NVf9V0YKze	Humans interact with the environment using a combination of perception - transforming sensory inputs from their environment into symbols, and cognition - mapping symbols to knowledge about the environment for supporting abstraction, reasoning by analogy, and long-term planning. Human perception-inspired machine perception, in the context of AI, refers to large-scale pattern recognition from raw data using neural networks trained using self-supervised learning objectives such as next-word prediction or object recognition. On the other hand, machine cognition encompasses more complex computations, such as using knowledge of the environment to guide reasoning, analogy, and long-term planning. Humans can also control and explain their cognitive functions. This seems to require the retention of symbolic mappings from perception outputs to knowledge about their environment. For example, humans can follow and explain the guidelines and safety constraints driving their decision-making in safety-critical applications such as healthcare, criminal justice, and autonomous driving. This article introduces the rapidly

	(open)	emerging paradigm of Neurosymbolic AI combines neural networks and knowledge-guided symbolic approaches to create more capable and flexible AI systems. These systems have immense potential to advance both algorithm-level (e.g., abstraction, analogy, reasoning) and application-level (e.g., explainable and safety-constrained decision-making) capabilities of AI systems.
2023-02	**ChatGPT Is a Blurry JPEG of the Web** Explainer at The New Yorker by Ted Chiang https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web	"OpenAI's chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?"
2023-02	**The Impact of AI on Developer Productivity: Evidence from GitHub Copilot** Sida Peng, Eirini Kalliamvakou, Peter Cihon, Mert Demirer https://arxiv.org/abs/2302.06590	Generative AI tools hold promise to increase human productivity. This paper presents results from a controlled experiment with GitHub Copilot, an AI pair programmer. Recruited software developers were asked to implement an HTTP server in JavaScript as quickly as possible. The treatment group, with access to the AI pair programmer, completed the task 55.8% faster than the control group. Observed heterogenous effects show promise for AI pair programmers to help people transition into software development careers.
2023-02	**What Google Should Really Be Worried About** Gary Marcus Substack Post 2023-02-12 https://garymarcus.substack.com/p/what-google-should-really-be-worried	"How sewers of lies could spell the end of web search"
2023-03	**GPT-4 Technical Report** Many authors Latest version 2024-03-04 https://arxiv.org/abs/2303.08774	🦉 **Introduces GTP-4. Unlike the GPT-3 paper, OpenAI does not disclose:** • **The number of parameters in GPT-4.** • **Details on the architecture or training dataset.** • **The training cost and exact hardware specifications.** We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.
2023-03	**Sparks of Artificial General Intelligence: Early experiments with GPT-4** Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke et al. https://arxiv.org/abs/2303.12712 Latest version: 2023-04-13	Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4, was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that moves beyond next-word prediction. We conclude with reflections on societal influences of the recent technological leap and future research directions.
2023-03	**The Unpredictable Abilities Emerging From Large AI Models** Popular Science explainer at Quanta Magazine by Stephen Ornes https://www.quantamagazine.org/the-unpredictable-abilities-emerging-from-large-ai-models-20230316/	"Large language models like ChatGPT are now big enough that they've started to display startling, unpredictable behaviors."
2023-03	**The AI Tech-Stack Model** Rua-Huan Tsaih, Hsin-Lu Chang, Chih-Chun Hsu, and David C. Yen Appears in: "Communications of the ACM", March 2023 https://cacm.acm.org/research/the-ai-tech-stack-model/	"Management and technology challenges of AI-enabled application projects."
2023-04	**Are Emergent Abilities of Large Language Models a Mirage?** Rylan Schaeffer, Brando Miranda, Sanmi Koyejo https://arxiv.org/abs/2304.15004	**"**Recent work claims that large language models display emergent abilities, abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their sharpness, transitioning seemingly instantaneously from not present to present, and their unpredictability, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: that for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in model behavior with scale. Specifically, nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous predictable changes in model performance..."
2023-05	**Google 'We Have No Moat, And Neither Does OpenAI'** By Dylan Patel and Afzal Ahmad	"Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI"

	https://semianalysis.com/2023/05/04/google-we-have-no-moat-and-neither/	
2023-05	**From Stochastic Parrots to Intelligent Assistants - The Secrets of Data and Human Interventions** Usama M. Fayyad Appears in "IEEE Intelligent Systems" May-June 2023 https://ieeexplore.ieee.org/document/10148666 (paywalled) https://www.computer.org/csdl/magazine/ex/2023/03/10148666/1NVfbJYtIR2 (open)	Generative AI is all the rage nowadays—primarily driven by ChatGPT capturing the public imagination and attracting hundreds of millions of users in record time, reaching 100 million users in two months. However, there is much ambiguity from the providers about the technology, the methodology, and the way OpenAI makes it work. This compounds the mystique and speculation. I focus on what we know, with a particular emphasis on the aspects that the makers of ChatGPT avoid discussing with the public—namely, the underlying dependence on much manual intervention in training data curation, data labeling, operational interventions by humans, and reinforcement learning. Unfortunately, despite the criticality of these issues to the scientific community, they are hardly discussed. In this article, I attempt to address some of the issues in the hope of stimulating further studies of these less glorified but critical topics.
2023-05	**Why an Octopus-like Creature Has Come to Symbolize the State of A.I.** Explainer at New York Times by Kevin Roose https://www.nytimes.com/2023/05/30/technology/shoggoth-meme-ai.html	"The Shoggoth, a character from a science fiction story, captures the essential weirdness of the A.I. moment."
2023-05	**Tiny Language Models Come of Age** Popular Science explainer at Quanta Magazine by Ben Brubaker https://www.quantamagazine.org/tiny-language-models-thrive-with-gpt-4-as-a-teacher-20231005/	"To better understand how neural networks learn to simulate writing, researchers trained simpler versions on synthetic children's stories."
2023-05	**Chatbots Don't Know What Stuff Isn't** Popular Science explainer at Quanta Magazine by Max G. Levy https://www.quantamagazine.org/ai-like-chatgpt-are-no-good-at-not-20230512/	"Today's language models are more sophisticated than ever, but they still struggle with the concept of negation. That's unlikely to change anytime soon."
2023-05	**ChatGPT's Astonishing Fabrications About Percy Ludgate** Brian Randell, Brian Coghlan Appears in: "IEEE Annals of the History of Computing", April-June 2023 https://ieeexplore.ieee.org/document/10148832 (paywalled) https://www.computer.org/csdl/magazine/an/2023/02/10148832/1NVeQNpkqnS (open) See also: **ChatGPT's Astonishing Fabrications About Percy Ludgate** Brian Coghlan, Brian Randell, Noel O'Boyle https://treasures.scss.tcd.ie/miscellany/TCD-SCSS-X.20121208.002/ChatGPTs-AstonishingFabrications-aboutPercyLudgate-CoghlanRandellOBoyle-20230424-1434.pdf	Since its release in November 2022, OpenAI's artificial intelligence (AI) chatbot ChatGPT has aroused great interest because of its very impressive ability to provide well-formulated and detailed natural language responses to queries about a huge variety of topics. These responses are based on an immense set of training data, obtained from the Internet in 2021, and on information gained from interactions with its users. However, ChatGPT's users soon found that the answers they received to their queries were not always trustworthy. Indeed, OpenAI itself lists as one of ChatGPT's limitations that it "sometimes writes plausible-sounding but incorrect or nonsensical answers, [i.e.,] confident responses that cannot be grounded in any of its training data." (The term "hallucination" has come into use by the AI community for such responses, which are not unique to ChatGPT.) *This should properly be called a "confabulation"*
2023-05	**Faith and Fate: Limits of Transformers on Compositionality** Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang et al. https://arxiv.org/abs/2305.18654 Latest version: 2023-10-31	Transformer large language models (LLMs) have sparked admiration for their exceptional performance on tasks that demand intricate multi-step reasoning. Yet, these models simultaneously show failures on surprisingly trivial problems. This begs the question: Are these errors incidental, or do they signal more substantial limitations? In an attempt to demystify transformer LLMs, we investigate the limits of these models across three representative compositional tasks -- multi-digit multiplication, logic grid puzzles, and a classic dynamic programming problem. These tasks require breaking problems down into sub-steps and synthesizing these steps into a precise answer. We formulate compositional tasks as computation graphs to systematically quantify the level of complexity, and break down reasoning steps into intermediate sub-procedures. Our empirical findings suggest that transformer LLMs solve compositional tasks by reducing multi-step compositional reasoning into linearized subgraph matching, without necessarily developing systematic problem-solving skills. To round off our empirical study, we provide theoretical arguments on abstract multi-step reasoning problems that highlight how autoregressive generations' performance can rapidly decay with increased task complexity.
2023-07	**Quaternion Convolutional Neural Networks: Current Advances and Future Directions** Gerardo Altamirano-Gomez, Carlos Gershenson https://arxiv.org/abs/2307.08663	Since their first applications, Convolutional Neural Networks (CNNs) have solved problems that have advanced the state-of-the-art in several domains. CNNs represent information using real numbers. Despite encouraging results, theoretical analysis shows that representations such as hyper-complex numbers can achieve richer representational capacities than real numbers, and that Hamilton products can capture intrinsic interchannel relationships. Moreover, in the last few years, experimental research has shown that Quaternion-Valued CNNs (QCNNs) can achieve similar performance with fewer parameters than their real-valued counterparts. This paper condenses research in the development of QCNNs from its very beginnings. We propose a conceptual organization of current trends and analyze the main building blocks used in the design of QCNN models. Based on this conceptual organization, we propose future directions of research.
2023-08	**Efficient Implementation of a Fully Analog Neural Network on a Reconfigurable Platform** Afolabi Ige, Jennifer Hasler Appears in: 2023 IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS) https://ieeexplore.ieee.org/document/10405875	This paper investigates the potential of floating gate field-effect transistors (FETs) as primitives for subthreshold computation in analog neural networks. By leveraging the inherent properties of these transistors, we demonstrate their suitability for constructing neural network activation functions, such as sigmoid and rectified linear units (ReLUs), as well as winner-take-all (WTA) circuits for softmax activation. Our end-to-end analog implementation successfully classifies the concentric circles problem, illustrating the advantages of maintaining an analog signal chain throughout the process.

2023-09	**Humble AI** Bran Knowles, Jason d'Cruz, John T. Richards, Kush R. Varsnhney Appears in: Communications of the ACM, 2023-09 https://dl.acm.org/doi/10.1145/3587035	An effort to bring artificial intelligence into better alignment with our moral aims and finally realize the vision of superior decision making through AI.
2023-10	**Does GPT-4 pass the Turing test?** Cameron R. Jones, Benjamin K. Bergen https://arxiv.org/abs/2310.20216	We evaluated GPT-4 in a public online Turing test. The best-performing GPT-4 prompt passed in 49.7% of games, outperforming ELIZA (22%) and GPT-3.5 (20%), but falling short of the baseline set by human participants (66%). Participants' decisions were based mainly on linguistic style (35%) and socioemotional traits (27%), supporting the idea that intelligence, narrowly conceived, is not sufficient to pass the Turing test. Participant knowledge about LLMs and number of games played positively correlated with accuracy in detecting AI, suggesting learning and practice as possible strategies to mitigate deception. Despite known limitations as a test of intelligence, we argue that the Turing test continues to be relevant as an assessment of naturalistic communication and deception. AI models with the ability to masquerade as humans could have widespread societal consequences, and we analyse the effectiveness of different strategies and criteria for judging humanlikeness.
2023-11	**Whom to Trust, How and Why: Untangling AI Ethics Principles, Trustworthiness and Trust** Andreas Duenser, David M. Douglas Appears in: "IEEE Intelligent Systems", Nov/Dec 2023 https://ieeexplore.ieee.org/document/10273868 https://arxiv.org/abs/2309.10318	We present an overview of the literature on trust in AI and AI trustworthiness and argue for the need to distinguish these concepts more clearly and to gather more empirically evidence on what contributes to people s trusting behaviours. We discuss that trust in AI involves not only reliance on the system itself, but also trust in the developers of the AI system. AI ethics principles such as explainability and transparency are often assumed to promote user trust, but empirical evidence of how such features actually affect how users perceive the system s trustworthiness is not as abundance or not that clear. AI systems should be recognised as socio-technical systems, where the people involved in designing, developing, deploying, and using the system are as important as the system for determining whether it is trustworthy. Without recognising these nuances, trust in AI and trustworthy AI risk becoming nebulous terms for any desirable feature for AI systems.
2023-11	**Seven Pillars for the Future of Artificial Intelligence** Erik Cambria, Rui Mao, Melvin Chen, Zhaoxia Wang, Seng-Beng Ho Appears in: "IEEE Intelligent Systems", Nov/Dec 2023 https://sentic.net/seven-pillars-for-the-future-of-artificial-intelligence.pdf https://www.computer.org/csdl/magazine/ex/2023/06/10352155/1SI3W47ymw8	*Coins the term "pareidoliac intelligence"* In recent years, AI research has showcased tremendous potential to impact positively humanity and society. Although AI frequently outperforms humans in tasks related to classification and pattern recognition, it continues to face challenges when dealing with complex tasks such as intuitive decision-making, sense disambiguation, sarcasm detection, and narrative understanding, as these require advanced kinds of reasoning, e.g., commonsense reasoning and causal reasoning, which have not been emulated satisfactorily yet. To address these shortcomings, we propose seven pillars that we believe represent the key hallmark features for the future of AI, namely: • Multidisciplinarity • Task Decomposition • Parallel Analogy • Symbol Grounding • Similarity Measure • Intention Awareness • Trustworthiness.
2023-12	**OpenAI Preparedness Framework (Beta)** https://cdn.openai.com/openai-preparedness-framework-beta.pdf See also: 2024-01-19: https://www.safer-ai.org/post/is-openais-preparedness-framework-better-than-its-competitors-responsible-scaling-policies-a-comparative-analysis	"Our analysis is that OpenAI's preparedness framework deals with accidental risks better than Anthropic's RSPs thanks to the more frequent assessments and the greater emphasis on risk identification and unknown unknowns. On the other hand, we believe that it deals less well with short-term catastrophic misuse risks, mostly due to the lack of details in the commitment of OpenAI to significant infosecurity measures which raises questions about how effectively they can prevent the leak of models with dangerous capabilities. One of the largest risks that 2024 or 2025 models might present is one where a system released publicly could increase catastrophic risks substantially when misused by malicious actors. Cybersecurity and infosecurity is the most important component to prevent this kind of disaster. This relative positioning of OpenAI towards accidental risks, compared with Anthropic's positioning towards misuse, reflects the respective worries of the CEOs of companies. Sam Altman, the CEO of OpenAI, is notably concerned about accidental extinction risks and has shared few public concerns about misuse risks. On the other hand, Dario Amodei, the CEO of Anthropic, has shared fewer concerns about accidental risks, but significantly more concerns about misuse risks."
2024		
2024-01	**Behind OpenAI CEO Dismissal: An Ethical Dilemma And A New AI Revolution** Ahmed El-Deeb, Amazon Appears in: ACM SIGSOFT Software Engineering Newsletter, Jan 2024	A very short note on the typical "business event"
2024-01	**New Theory Suggests Chatbots Can Understand Text** Popular Science explainer at Quanta Magazine by Anil Ananthaswamy https://www.quantamagazine.org/new-theory-suggests-chatbots-can-understand-text-20240122/	"Far from being 'stochastic parrots,' the biggest large language models seem to learn enough skills to understand the words they're processing."
2024-01	**Exploring Large Language Model based Intelligent Agents: Definitions, Methods, and Prospects** Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, Xiuqiang He https://arxiv.org/abs/2401.03428	Intelligent agents stand out as a potential path toward artificial general intelligence (AGI). Thus, researchers have dedicated significant effort to diverse implementations for them. Benefiting from recent progress in large language models (LLMs), LLM-based agents that use universal natural language as an interface exhibit robust generalization capabilities across various applications -- from serving as autonomous general-purpose task assistants to applications in coding, social, and economic domains, LLM-based agents offer extensive exploration opportunities. This paper surveys current research to provide an in-depth overview of LLM-based intelligent agents within single-agent and multi-agent systems. It covers their definitions, research frameworks, and foundational components such as their composition, cognitive and planning methods, tool utilization, and responses to environmental feedback. We also delve into the mechanisms of deploying LLM-based

		agents in multi-agent systems, including multi-role collaboration, message passing, and strategies to alleviate communication issues between agents. The discussions also shed light on popular datasets and application scenarios. We conclude by envisioning prospects for LLM-based agents, considering the evolving landscape of AI and natural language processing.
2024-01	**DeepSeek LLM: Scaling Open-Source Language Models with Longtermism** DeepSeek-AI https://arxiv.org/abs/2401.02954	The rapid development of open-source large language models (LLMs) has been truly remarkable. However, the scaling law described in previous literature presents varying conclusions, which casts a dark cloud over scaling LLMs. We delve into the study of scaling laws and present our distinctive findings that facilitate scaling of large scale models in two commonly used open-source configurations, 7B and 67B. Guided by the scaling laws, we introduce DeepSeek LLM, a project dedicated to advancing open-source language models with a long-term perspective. To support the pre-training phase, we have developed a dataset that currently consists of 2 trillion tokens and is continuously expanding. We further conduct supervised fine-tuning (SFT) and Direct Preference Optimization (DPO) on DeepSeek LLM Base models, resulting in the creation of DeepSeek Chat models. Our evaluation results demonstrate that DeepSeek LLM 67B surpasses LLaMA-2 70B on various benchmarks, particularly in the domains of code, mathematics, and reasoning. Furthermore, open-ended evaluations reveal that DeepSeek LLM 67B Chat exhibits superior performance compared to GPT-3.5.
2024-02	**On Limitations of the Transformer Architecture** Binghui Peng, Srini Narayanan, Christos Papadimitriou https://arxiv.org/abs/2402.08164	What are the root causes of hallucinations in large language models (LLMs)? We use Communication Complexity to prove that the Transformer layer is incapable of composing functions (e.g., identify a grandparent of a person in a genealogy) if the domains of the functions are large enough; we show through examples that this inability is already empirically present when the domains are quite small. We also point out that several mathematical tasks that are at the core of the so-called compositional tasks thought to be hard for LLMs are unlikely to be solvable by Transformers, for large enough instances and assuming that certain well accepted conjectures in the field of Computational Complexity are true.
2024-02	**How Quickly Do Large Language Models Learn Unexpected Skills?** Popular Science explainer at Quanta Magazine by Stephen Ornes https://www.quantamagazine.org/how-quickly-do-large-language-models-learn-unexpected-skills-20240213/	"A new study suggests that so-called emergent abilities actually develop gradually and predictably, depending on how you measure them."
2024-02	**Measuring GitHub Copilot's Impact on Productivity** Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian Appears in "Communications of the ACM", March 2024 https://cacm.acm.org/research/measuring-github-copilots-impact-on-productivity/	"A case study asks Copilot users about the tool's impact on their productivity, and seeks to find their perceptions mirrored in user data."
2024-03	**Can Machines Be in Language?** Peter J. Denning and B. Scot Rousse Appears in "Communications of the ACM", March 2024 https://cacm.acm.org/opinion/can-machines-be-in-language/ https://dl.acm.org/doi/10.1145/3637629	"Large language models brought language to machines. Machines are not up to the challenge"
2024-03	**Can ChatGPT Learn Chinese or Swahili?** Neil Savage Appears in "Communications of the ACM", May 2024 https://cacm.acm.org/news/can-chatgpt-learn-chinese-or-swahili/ https://dl.acm.org/doi/abs/10.1145/3640351	"Considering how large language models might act differently if trained in different languages."
2024-03	**Assured Autonomy, Artificial Intelligence, and Machine Learning: A Roundtable Discussion** • Phil Laplante , IEEE Fellow • Joanna F. DeFranco , The Pennsylvania State University • Rick Kuhn , IEEE Fellow Appears in: "IEEE Computer", March 2024 https://www.computer.org/csdl/magazine/co/2024/03/10461703/1V5LYnygCOc	This report summarizes a roundtable panel discussion held at the Second Annual IEEE Workshop on Assured Autonomy, AI, and Machine Learning. Eight expert panelists discussed ways to ensure that artificial intelligence and machine learning systems are safe. Roundtable Panelists: • *Jaganmohan Chandrasekaran* is a postdoctoral associate at the Intelligent Systems Division, Virginia Tech National Security Institute. • *Darren Cofer* is a principal fellow at Collins Aerospace. • *Junhua Ding* is the Reinburg Endowed Professor in Data Science and the director of the Graduate Data Science Program, Department of Information Science, University of North Texas. • *Carl Elks* is an associate professor of electrical and computer engineering at Virginia Commonwealth University. • *Cody Fleming* is an associate professor of mechanical engineering at Iowa State University. • *Alwyn Goodloe* is a research computer engineer at NASA Langley Research Center. • *Erin Lanus* is a research assistant professor at the Intelligent Systems Division, Virginia Tech National Security Institute. • *Adam Porter* is a professor at the University of Maryland and the director of the Fraunhofer USA Center Mid-Atlantic.
2024-03	**Grounding from an AI and Cognitive Science Lens** Goonmeet Bajaj, Valerie L. Shalin, Srinivasan Parthasarathy, Amit Sheth Appears in: "IEEE Intelligent Systems", March/April 2024 https://ieeexplore.ieee.org/document/10510670	Grounding is a challenging problem, requiring a formal definition and different levels of abstraction. This article explores grounding from both cognitive science and machine learning perspectives. It identifies the subtleties of grounding, its significance for collaborative agents, and similarities and differences in grounding approaches in both communities. The article examines the potential of neurosymbolic approaches tailored for grounding tasks, showcasing how they can more comprehensively address grounding. Finally, we discuss areas for further exploration and development in grounding.

	https://arxiv.org/abs/2402.13290	Natural Language Processing / Computer Vision: static vs dynamic Static • Connecting concepts to known data • Axiomatic common ground Dynamic • Iterative method to establish common ground • Communication between agent(s) to clarify and correct Cognitive science: cognitivist vs. enactivist Cognitivist • Connecting atomic representations to external entities • Compositional representations Enactivist • Establishing physical connection in an environment • Embodies grounding through sensory input and feedback
2024-05	**Limits of Deep Learning: Sequence Modeling through the Lens of Complexity Theory** Nikola Zubić, Federico Soldá, Aurelio Sulser, Davide Scaramuzza https://arxiv.org/abs/2405.16674 Latest version: 2024-10-04	Despite their successes, deep learning models struggle with tasks requiring complex reasoning and function composition. We present a theoretical and empirical investigation into the limitations of Structured State Space Models (SSMs) and Transformers in such tasks. We prove that one-layer SSMs cannot efficiently perform function composition over large domains without impractically large state sizes, and even with Chain-of-Thought prompting, they require a number of steps that scale unfavorably with the complexity of the function composition. Multi-layer SSMs are constrained by log-space computational capacity, limiting their reasoning abilities. Our experiments corroborate these theoretical findings. Evaluating models on tasks including various function composition settings, multi-digit multiplication, dynamic programming, and Einstein's puzzle, we find significant performance degradation even with advanced prompting techniques. Models often resort to shortcuts, leading to compounding errors. These findings highlight fundamental barriers within current deep learning architectures rooted in their computational capacities. We underscore the need for innovative solutions to transcend these constraints and achieve reliable multi-step reasoning and compositional task-solving, which is critical for advancing toward general artificial intelligence.
2024-05	**DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model** DeepSeek-AI Latest version 2024-06-19 https://arxiv.org/abs/2405.04434	We present DeepSeek-V2, a strong Mixture-of-Experts (MoE) language model characterized by economical training and efficient inference. It comprises 236B total parameters, of which 21B are activated for each token, and supports a context length of 128K tokens. DeepSeek-V2 adopts innovative architectures including Multi-head Latent Attention (MLA) and DeepSeekMoE. MLA guarantees efficient inference through significantly compressing the Key-Value (KV) cache into a latent vector, while DeepSeekMoE enables training strong models at an economical cost through sparse computation. Compared with DeepSeek 67B, DeepSeek-V2 achieves significantly stronger performance, and meanwhile saves 42.5% of training costs, reduces the KV cache by 93.3%, and boosts the maximum generation throughput to 5.76 times. We pretrain DeepSeek-V2 on a high-quality and multi-source corpus consisting of 8.1T tokens, and further perform Supervised Fine-Tuning (SFT) and Reinforcement Learning (RL) to fully unlock its potential. Evaluation results show that, even with only 21B activated parameters, DeepSeek-V2 and its chat versions still achieve top-tier performance among open-source models.
2024-06 📘	**A Review of Pulse-Coupled Neural Network Applications in Computer Vision and Image Processing** Nurul Rafi and Pablo Rivas https://arxiv.org/pdf/2406.00239	Research in neural models inspired by mammal's visual cortex has led to many spiking neural networks such as pulse-coupled neural networks (PCNNs). These models are oscillating, spatio-temporal models stimulated with images to produce several time-based responses. This paper reviews PCNN's state of the art, covering its mathematical formulation, variants, and other simplifications found in the literature. We present several applications in which PCNN architectures have successfully addressed some fundamental image processing and computer vision challenges, including image segmentation, edge detection, medical imaging, image fusion, image compression, object recognition, and remote sensing. Results achieved in these applications suggest that the PCNN architecture generates useful perceptual information relevant to a wide variety of computer vision tasks.
2024-06 📘	**A comprehensive survey on Kolmogorov Arnold Networks (KAN)** Tianrui Ji, Yuntian Hou, Di Zhang https://arxiv.org/abs/2407.11075 Latest version is dated 2025-01-28 See also: https://www.quantamagazine.org/novel-architecture-makes-neural-networks-more-understandable-20240911/	Through this comprehensive survey of Kolmogorov-Arnold Networks(KAN), we have gained a thorough understanding of its theoretical foundation, architectural design, application scenarios, and current research progress. KAN, with its unique architecture and flexible activation functions, excels in handling complex data patterns and nonlinear relationships, demonstrating wide-ranging application potential. While challenges remain, KAN is poised to pave the way for innovative solutions in various fields, potentially revolutionizing how we approach complex computational problems.
2024-06	**Automatic Programming vs. Artificial Intelligence** James Noble Appears in. "AIware 2024: Proceedings of the 1st ACM International Conference on AI-Powered Software" https://dl.acm.org/doi/10.1145/3664646.3664775	Ever since we began programming in the 1950s, there have been two diametrically opposed tendencies within computer science and software engineering: on the left side of the Glorious Throne of Alan Turing, the tendency to perfect the Art of Computer Programming, and on the right side, the tendency to end it. These tendencies can be seen from the Manchester Mark I's "autocode" removing the need for programmers shortly after WW2; COBOL being a language that could be "read by the management"; to contemporary "no-code" development environments; and the idea that large language models herald "The End of Programming". This vision paper looks at what AI will not change about software systems, and the people who must use them, and necessarily must build them. Rather than neglecting 50 years of history, theory, and practice, and assuming programming can, will, and should be ended by AI, we speculate on how AI has, already does, and will continue to perfect one of the peak activities of being human: programming.
2024-07	**The Llama 3 Herd of Models** Llama Team, AI @ Meta https://arxiv.org/abs/2407.21783 Latest version 2024-11-23	🦉 Modern artificial intelligence (AI) systems are powered by foundation models. This paper presents a new set of foundation models, called Llama 3. It is a herd of language models that natively support multilinguality, coding, reasoning, and tool usage. Our largest model is a dense Transformer with 405B parameters and a context window of up to 128K tokens. This paper presents an extensive empirical evaluation of Llama 3. We find that Llama 3 delivers comparable quality to leading language models such as GPT-4 on a plethora of tasks. We publicly release Llama 3, including pre-trained and post-trained versions of the 405B parameter language model and our Llama Guard 3 model for input and output safety. The paper also presents the results of experiments in which we integrate image, video, and speech

		capabilities into Llama 3 via a compositional approach. We observe this approach performs competitively with the state-of-the-art on image, video, and speech recognition tasks. The resulting models are not yet being broadly released as they are still under development.
2024-08	**Debate May Help AI Models Converge on Truth** Popular Science explainer at Quanta Magazine by Stephen Ornes https://www.quantamagazine.org/debate-may-help-ai-models-converge-on-truth-20241108/	"Letting AI systems argue with each other may help expose when a large language model has made mistakes."
2024-08	**Fire-Flyer AI-HPC: A Cost-Effective Software-Hardware Co-Design for Deep Learning** DeepSeek-AI, Beijing, China Open: https://arxiv.org/abs/2408.14158 Paywalled: https://ieeexplore.ieee.org/document/10793193 Published in: "SC24: International Conference for High Performance Computing, Networking, Storage and Analysis", 17-22 November 2024	The rapid progress in Deep Learning (DL) and Large Language Models (LLMs) has exponentially increased demands of computational power and bandwidth. This, combined with the high costs of faster computing chips and interconnects, has significantly inflated High Performance Computing (HPC) construction costs. To address these challenges, we introduce the Fire-Flyer AI-HPC architecture, a synergistic hardware-software co-design framework and its best practices. (...)
2024-08 📘	**Toward Large-scale Spiking Neural Networks: A Comprehensive Survey and Future Directions** Yangfan Hu, Qian Zheng, Guoqi Li, Huajin Tang, Gang Pan https://arxiv.org/abs/2409.02111	Deep learning has revolutionized artificial intelligence (AI), achieving remarkable progress in fields such as computer vision, speech recognition, and natural language processing. Moreover, the recent success of large language models (LLMs) has fueled a surge in research on large-scale neural networks. However, the escalating demand for computing resources and energy consumption has prompted the search for energy-efficient alternatives. Inspired by the human brain, spiking neural networks (SNNs) promise energy-efficient computation with event-driven spikes. To provide future directions toward building energy-efficient large SNN models, we present a survey of existing methods for developing deep spiking neural networks, with a focus on emerging Spiking Transformers. Our main contributions are as follows: (1) an overview of learning methods for deep spiking neural networks, categorized by ANN-to-SNN conversion and direct training with surrogate gradients; (2) an overview of network architectures for deep spiking neural networks, categorized by deep convolutional neural networks (DCNNs) and Transformer architecture; and (3) a comprehensive comparison of state-of-the-art deep SNNs with a focus on emerging Spiking Transformers. We then further discuss and outline future directions toward large-scale SNNs.
2024-10 📘	**A Comprehensive Overview of Large Language Models** Earliest version 2023-07-12 Latest version 2024-10 Humza Naveeda, Asad Ullah Khana, Shi Qiub, Muhammad Saqibc, et al. https://arxiv.org/abs/2307.06435	Large Language Models (LLMs) have recently demonstrated remarkable capabilities in natural language processing tasks and beyond. This success of LLMs has led to a large influx of research contributions in this direction. These works encompass diverse topics such as architectural innovations, better training strategies, context length improvements, fine-tuning, multi-modal LLMs, robotics, datasets, benchmarking, efficiency, and more. With the rapid development of techniques and regular breakthroughs in LLM research, it has become considerably challenging to perceive the bigger picture of the advances in this direction. Considering the rapidly emerging plethora of literature on LLMs, it is imperative that the research community is able to benefit from a concise yet comprehensive overview of the recent developments in this field. This article provides an overview of the existing literature on a broad range of LLM-related concepts. Our self-contained comprehensive overview of LLMs discusses relevant background concepts along with covering the advanced topics at the frontier of research in LLMs. This review article is intended to not only provide a systematic survey but also a quick comprehensive reference for the researchers and practitioners to draw insights from extensive informative summaries of the existing works to advance the LLM research.
2025-10	**An Evolved Universal Transformer Memory** Edoardo Cetin, Qi Sun, Tianyu Zhao, Yujin Tang https://arxiv.org/abs/2410.13166	Prior methods propose to offset the escalating costs of modern foundation models by dropping specific parts of their contexts with hand-designed rules, while attempting to preserve their original performance. We overcome this trade-off with Neural Attention Memory Models (NAMMs), introducing a learned network for memory management that improves both the performance and efficiency of transformers. We evolve NAMMs atop pre-trained transformers to provide different latent contexts focusing on the most relevant information for individual layers and attention heads. NAMMs are universally applicable to any model using self-attention as they condition exclusively on the values in the produced attention matrices. Learning NAMMs on a small set of problems, we achieve substantial performance improvements across multiple long-context benchmarks while cutting the model's input contexts up to a fraction of the original sizes. We show the generality of our conditioning enables zero-shot transfer of NAMMs trained only on language to entirely new transformer architectures even across input modalities, with their benefits carrying over to vision and reinforcement learning.
2024-10	**Reevaluating Google's Reinforcement Learning for IC Macro Placement** Igor L. Markov https://cacm.acm.org/research/reevaluating-googles-reinforcement-learning-for-ic-macro-placement/ Appears in: "Communications of the ACM" November 2024	A 2021 paper in *Nature* by Mirhoseini, Goldie, et al. (1) about the use of reinforcement learning (RL) in the physical design of silicon chips raised eyebrows, drew critical media coverage, and stirred up controversy due to poorly documented claims. The paper, authored by Google researchers, withheld critical methodological steps, and most inputs needed to reproduce its results. Our meta-analysis shows how two separate evaluations filled in the gaps and demonstrated that Google RL lags behind human chip designers, a well-known algorithm (simulated annealing), and generally available commercial software, while also being slower. Crosschecked data indicates that the integrity of the *Nature* paper is substantially undermined, owing to errors in conduct, analysis, and reporting. Before publishing, Google rebuffed internal allegations of fraud which still stand. We note policy implications. (1) Mazyavkina, N., Sviridov, S., Ivanov, S., and Burnaev, E. Reinforcement learning for combinatorial optimization: A survey. *Computers and Operations Research 134* (2021), 105400; https://bit.ly/4dsfbKV
2024-11	**AI Should Challenge, Not Obey** Advait Sarkar, Microsoft Research Appears in "Communications of the ACM", October 2024 https://cacm.acm.org/opinion/ai-should-challenge-not-obey/ https://arxiv.org/abs/2411.02263	"Let's transform our robot secretaries into Socratic gadflies."
2024-12	**Is It Possible to Truly Understand Performance in LLMs?**	Seeking to understand when, and how, new skills and capabilities emerge in LLMs.

	Samuel Greengard Appears in "Communications of the ACM", December 2024 https://cacm.acm.org/news/is-it-possible-to-truly-understand-performance-in-llms/	
2024-12	**An AI Learning Hierarchy** Peter J. Denning, Ted G. Lewis https://cacm.acm.org/opinion/an-ai-learning-hierarchy/ Appears in: "Communications of the ACM", December 2024	A hierarchy of AI machines organized by their learning power shows their limits and the possibility that humans are at risk of machine subjugation well before AI utopia can come. Level Category of Machines 0 Basic automation 1 Rule-based systems 2 Supervised learning 3 Unsupervised learning 4 Generative AI 5 Reinforcement learning AI 6 Human-machine interaction AI 7 Aspirational AI See also: "Intelligence may not be computable", by the same authors, 2019-12.
2024-12	**DeepSeek-V3 Technical Report** DeepSeek-AI https://arxiv.org/abs/2412.19437	We present DeepSeek-V3, a strong Mixture-of-Experts (MoE) language model with 671B total parameters with 37B activated for each token. To achieve efficient inference and cost-effective training, DeepSeek-V3 adopts Multi-head Latent Attention (MLA) and DeepSeekMoE architectures, which were thoroughly validated in DeepSeek-V2. Furthermore, DeepSeek-V3 pioneers an auxiliary-loss-free strategy for load balancing and sets a multi-token prediction training objective for stronger performance. (...)
2024-12	**The EU AI Act and the Wager on Trustworthy AI** Appears in: "Communications of the ACM", December 2024 https://cacm.acm.org/research/the-eu-ai-act-and-the-wager-on-trustworthy-ai	As the impact of AI is difficult to assess by a single group, policymakers should prioritize societal and environmental well being and seek advice from interdisciplinary groups focusing on ethical aspects, responsibility, and transparency in the development of algorithms.
2024-12	**Prompting Considered Harmful** Meredith Ringel Morris Appears in: "Communications of the ACM", December 2024 https://cacm.acm.org/opinion/prompting-considered-harmful/	As systems graduate from labs to the open world, moving beyond prompting is central to ensuring that AI is useful, usable, and safe for end users as well as experts.
2024-12	**States as Strings as Strategies: Steering Language Models with Game-Theoretic Solvers** Ian Gemp, Yoram Bachrach, Marc Lanctot, Roma Patel, Vibhavari Dasagi, Luke Marris, Georgios Piliouras, Karl Tuyls https://arxiv.org/abs/2402.01704	Mathematical models of interactions among rational agents have long been studied in game theory. However these interactions are often over a small set of discrete game actions which is very different from how humans communicate in natural language. To bridge this gap, we introduce a framework that allows equilibrium solvers to work over the space of natural language dialogue generated by large language models (LLMs). Specifically, by modelling the players, strategies and payoffs in a "game" of dialogue, we create a binding from natural language interactions to the conventional symbolic logic of game theory. Given this binding, we can ask existing game-theoretic algorithms to provide us with strategic solutions (e.g., what string an LLM should generate to maximize payoff in the face of strategic partners or opponents), giving us predictors of stable, rational conversational strategies. We focus on three domains that require different negotiation strategies: scheduling meetings, trading fruit and debate, and evaluate an LLM's generated language when guided by solvers. We see that LLMs that follow game-theory solvers result in dialogue generations that are less exploitable than the control (no guidance from solvers), and the language generated results in higher rewards, in all negotiation domains. We discuss future implications of this work, and how game-theoretic solvers that can leverage the expressivity of natural language can open up a new avenue of guiding language research.
2024-12	**The Emergence of Strategic Reasoning of Large Language Models** Dongwoo Lee, Gavin Kader https://arxiv.org/abs/2412.13013	As Large Language Models (LLMs) are increasingly used for a variety of complex and critical tasks, it is vital to assess their logical capabilities in strategic environments. This paper examines their ability in strategic reasoning -- the process of choosing an optimal course of action by predicting and adapting to other agents' behavior. Using six LLMs, we analyze responses from play in classical games from behavioral economics (p-Beauty Contest, 11-20 Money Request Game, and Guessing Game) and evaluate their performance through hierarchical models of reasoning (level-k theory and cognitive hierarchy theory). Our findings reveal that while LLMs show understanding of the games, the majority struggle with higher-order strategic reasoning. Although most LLMs did demonstrate learning ability with games involving repeated interactions, they still consistently fall short of the reasoning levels demonstrated by typical behavior from human subjects. The exception to these overall findings is with OpenAI's GPT-o1 -- specifically trained to solve complex reasoning tasks -- which consistently outperforms other LLMs and human subjects. These findings highlight the challenges and pathways in advancing LLMs toward robust strategic reasoning from the perspective of behavioral economics.
2024-12	**Titans: Learning to Memorize at Test Time** Ali Behrouz, Peilin Zhong, Vahab Mirrokni https://arxiv.org/abs/2501.00663v1	Over more than a decade there has been an extensive research effort on how to effectively utilize recurrent models and attention. While recurrent models aim to compress the data into a fixed-size memory (called hidden state), attention allows attending to the entire context window, capturing the direct dependencies of all tokens. This more accurate modeling of dependencies, however, comes with a quadratic cost, limiting the model to a fixed-length context. We present a new neural long-term memory module that learns to memorize historical context and helps attention to attend to the current context while utilizing long past information. We show that this neural memory has the advantage of fast parallelizable training while maintaining a fast inference. From a memory perspective, we argue that attention due to its limited context but accurate dependency modeling performs as a short-term memory, while neural memory due to its ability to memorize the data, acts as a long-term, more persistent, memory. Based on these two modules, we introduce a new family of architectures, called Titans, and present three variants to address how

		one can effectively incorporate memory into this architecture. Our experimental results on language modeling, common-sense reasoning, genomics, and time series tasks show that Titans are more effective than Transformers and recent modern linear recurrent models. They further can effectively scale to larger than 2M context window size with higher accuracy in needle-in-haystack tasks compared to baselines.
2024-12	**Memory Layers at Scale** Vincent-Pierre Berges, Barlas Oğuz, Daniel Haziza, Wen-tau Yih, Luke Zettlemoyer, Gargi Ghosh https://arxiv.org/abs/2412.09764	Memory layers use a trainable key-value lookup mechanism to add extra parameters to a model without increasing FLOPs. Conceptually, sparsely activated memory layers complement compute-heavy dense feed-forward layers, providing dedicated capacity to store and retrieve information cheaply. This work takes memory layers beyond proof-of-concept, proving their utility at contemporary scale. On downstream tasks, language models augmented with our improved memory layer outperform dense models with more than twice the computation budget, as well as mixture-of-expert models when matched for both compute and parameters. We find gains are especially pronounced for factual tasks. We provide a fully parallelizable memory layer implementation, demonstrating scaling laws with up to 128B memory parameters, pretrained to 1 trillion tokens, comparing to base models with up to 8B parameters.
2025		
2025-01	**Can AI Models Show Us How People Learn? Impossible Languages Point a Way.** https://www.quantamagazine.org/can-ai-models-show-us-how-people-learn-impossible-languages-point-a-way-20250113/ https://arxiv.org/abs/2401.06416 - Mission: Impossible Language Models Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, Christopher Potts	Popular Science explainer at Quanta Magazine by Ben Brubaker
2025-01	**Unlearning in Large Language Models: We Are Not There Yet** Alberto Blanco-Justicia, Josep Domingo-Ferrer, Najeeb Moharram Jebreel et al. Appears in: "IEEE Computer", January 2025 https://ieeexplore.ieee.org/document/10834279	The massive adoption of large language models has prompted concerns about how to align them with human ethics and the rule of law. Digital forgetting of undesirable knowledge via machine unlearning is a promising strategy we survey here.
2025-01	**Qwen2.5 Technical Report** Qwen Team https://arxiv.org/abs/2412.15115	In this report, we introduce Qwen2.5, a comprehensive series of large language models (LLMs) designed to meet diverse needs. Compared to previous iterations, Qwen 2.5 has been significantly improved during both the pre-training and post-training stages. In terms of pre-training, we have scaled the high-quality pre-training datasets from the previous 7 trillion tokens to 18 trillion tokens. This provides a strong foundation for common sense, expert knowledge, and reasoning capabilities. In terms of post-training, we implement intricate supervised finetuning with over 1 million samples, as well as multistage reinforcement learning. Post-training techniques enhance human preference, and notably improve long text generation, structural data analysis, and instruction following. To handle diverse and varied use cases effectively, we present Qwen2.5 LLM series in rich sizes. Open-weight offerings include base and instruction-tuned models, with quantized versions available.
2025-01	**People who frequently use ChatGPT for writing tasks are accurate and robust detectors of AI-generated text** Jenna Russell, Marzena Karpinska, Mohit Iyyer https://arxiv.org/abs/2501.15654	In this paper, we study how well humans can detect text generated by commercial LLMs (GPT-4o, Claude, o1). We hire annotators to read 300 non-fiction English articles, label them as either human-written or AI-generated, and provide paragraph-length explanations for their decisions. Our experiments show that annotators who frequently use LLMs for writing tasks excel at detecting AI-generated text, even without any specialized training or feedback. In fact, the majority vote among five such "expert" annotators misclassifies only 1 of 300 articles, significantly outperforming most commercial and open-source detectors we evaluated even in the presence of evasion tactics like paraphrasing and humanization. Qualitative analysis of the experts' free-form explanations shows that while they rely heavily on specific lexical clues ('AI vocabulary'), they also pick up on more complex phenomena within the text (e.g., formality, originality, clarity) that are challenging to assess for automatic detectors. We release our annotated dataset and code to spur future research into both human and automated detection of AI-generated text.
2025-01	**Kimi k1.5: Scaling Reinforcement Learning with LLMs** Kimi Team https://arxiv.org/abs/2501.12599	Language model pretraining with next token prediction has proved effective for scaling compute but is limited to the amount of available training data. Scaling reinforcement learning (RL) unlocks a new axis for the continued improvement of artificial intelligence, with the promise that large language models (LLMs) can scale their training data by learning to explore with rewards. However, prior published work has not produced competitive results. In light of this, we report on the training practice of Kimi k1.5, our latest multi-modal LLM trained with RL, including its RL training techniques, multi-modal data recipes, and infrastructure optimization. Long context scaling and improved policy optimization methods are key ingredients of our approach, which establishes a simplistic, effective RL framework without relying on more complex techniques such as Monte Carlo tree search, value functions, and process reward models.
2025-01	**Chatbot Software Begins to Face Fundamental Limitations** Popular Science explainer by Anil Ananthaswamy https://www.quantamagazine.org/chatbot-software-begins-to-face-fundamental-limitations-20250131	"Recent results show that large language models struggle with compositional tasks, suggesting a hard limit to their abilities."
2025-02	**Deep Research, Deep Bullshit, and the potential (model) collapse of science** Gary Marcus Substack post, 2025-02-03 https://garymarcus.substack.com/p/deep-research-deep-bullshit-and-the	"Sam Altman’s hype might just bite us all in the behind"