.

Jürgen Schmidhuber
Pronounce: You_again Shmidhoobuh
Technical Report IDSIA-24-24

AI Blog
Twitter: @SchmidhuberAI
7 Dec 2024 (v1), updated 31 July 2025 (v2)

A Nobel Prize for Plagiarism

Abstract. Sadly, the Nobel Prize in Physics 2024 for Hopfield & Hinton is a Nobel Prize for plagiarism. They republished methodologies for artificial neural networks developed in Ukraine and Japan by Ivakhnenko and Amari in the 1960s & 1970s, as well as other techniques, without citing the original papers. Even in later surveys, they didn't credit the original inventors (thus turning what may have been unintentional plagiarism into a deliberate form). None of the important algorithms for modern Artificial Intelligence were created by Hopfield & Hinton.

The field of machine intelligence is rife with plagiarism. I am not even talking about platform companies and other providers of Large Language Models (LLMs) that have been accused of stealing the work of numerous authors and artists on the web.[PLA1] No, I am referring to human AI researchers plagiarising other human AI researchers. I will illustrate the problem by analyzing the background of a particularly prominent recent award.

Modern AI is based on what's now called "deep learning" with artificial neural networks (NNs).[DL1][DLH] The Nobel Prize in Physics 2024 was awarded for *"foundational discoveries that enable machine learning with artificial NNs."*[Nob24a][NOBnat] Unfortunately, however, the awardees did not make any such foundational discoveries. Instead they republished methodologies developed in

Ukraine and Japan in the 1960s and 1970s, as well as other techniques, without citing the original inventors. See Sec. 1-4 below.

I am one of the persons cited by the Nobel Foundation in the *Scientific Background to the Nobel Prize in Physics 2024.*[Nob24a] The most cited neural networks and AIs all build on work done in my labs,[MOST] including the most cited AI paper of the 20th century.[LSTM1] I am also known for the most comprehensive surveys of modern AI and deep learning.[DL1][DLH]

I made a popular tweet on this prize, and followed up with a more detailed technical report.[NOB] This award directly rewards plagiarism and misattribution, which is part of a dark trend in AI that has been going on for some time. The Nobel Committee has now welcomed this malevolence into physics with open arms.

Is it now acceptable for me to direct young Ph.D. students to read old papers and rewrite and resubmit them as if they were their own works? Whatever the intention, this award says that, yes, that is perfectly fine.

Some people have lost their titles or jobs due to plagiarism, e.g., Harvard's former president.[PLAG7] But after this Nobel Prize, how can advisors now continue to tell their students that they should avoid plagiarism at all costs?

Of course, it is well known that plagiarism can be either "unintentional" or "intentional or reckless,"[PLAG1-6] and the more innocent of the two may very well be partially the case here. But science has a well-established way of dealing with "multiple discovery" and plagiarism—be it unintentional[PLAG1-6][CONN21] or not[FAKE2]—based on facts such as time stamps of publications and patents. The deontology of science requires that unintentional plagiarists correct their publications through errata and then credit the original sources properly in the future. The awardees didn't; instead the awardees kept collecting citations for inventions of other researchers.[DLP] This behaviour turns even *un*intentional plagiarism[PLAG1-6] into an *intentional* form.[FAKE2]

Since the first version of this report came out in 2024, the plagiarism allegations[NOB][DLP] have essentially remained unchallenged: the awardees have not tried to defend themselves. They can't—the facts are the facts.

1. Hopfield vs Amari and others

The Lenz-Ising recurrent architecture with neuron-like elements was published in 1925.[L20][I24,I25] In 1972, Shun-Ichi Amari made it adaptive such that it could learn to associate a finite number of input patterns with output patterns by changing its connection weights.[AMH1] The weights are correlations of the stored patterns. A stored pattern is recalled from similar patterns through the neural dynamics.

Amari's approach was republished much later by Hopfield who used the same basic equations but failed to cite the prior work.[AMH2][DLP]

Unfortunately, Amari is only briefly cited in the *Scientific Background to the Nobel Prize in Physics 2024.*[Nob24a] The Nobel Prize is about *“foundational discoveries that enable machine learning with artificial NNs.”* Amari obviously published this *"foundational discovery"* 10 years

before Hopfield, when compute was about 100 times more expensive. However, even Hopfield's much later survey of "Hopfield networks" (Scholarpedia, 2007)[AMH4] failed to cite Amari (1972).

Note that making slight modifications doesn't let you ignore the work that introduced the key innovation. These are just the elementary rules of scientific publishing. Hopfield's own contributions build on the prior work: an analysis of storage capacity and a suitable Lyapunov function for sequential neuron updates instead of Amari's parallel updates (in practice, this hardly makes a difference[AM24]) to show that the "Hopfield network" settles into an equilibrium in response to static input patterns. [AMH2] These sequential updates and equilibrium nets are mostly irrelevant in modern AI, which uses massively parallel neuron updates and focuses heavily on sequence processing.[DLH] Note that Amari (1972) already had a sequence-processing generalization of the "Hopfield network."[AMH1]

The Nobel Committee for Physics[Nob24a] briefly cites the important work of Nakano (1972),[NAK72] [DLH] a former collaborator of Amari. Nakano also had a recurrent associative memory, but it had ternary activation values and wasn't the "Hopfield network"—sometimes called the "Amari-Hopfield Network"[AMH3]— first published by Amari (1972).[AMH1][DLH]

Remarkably, Hopfield[AMH2] was aware of Amari: he cites Amari's *later* papers on the separate topic of self-organisation in NNs (1977, 1978), but ignores his crucial 1972 paper[AMH1][DLH] (as well as Nakano's paper). (See also Little's work (1974-1980)[AMH1b-d] on connecting the Lenz-Ising network (1920s)[L20][I24,I25] to learning NNs.)

In 1984, Hopfield published an analog version[HOP84][Nob24a] that failed to cite Cohen and Grossberg's earlier work (1983),[GRO83] which described a Lyapunov function for the "Additive Model," later called the "Hopfield model." This publication built on Grossberg's even earlier work (1978) on the Additive Model,[GRO78] which introduced a Lyapunov functional to help prove convergence. See also Grossberg's overviews.[GRO20][GRO21] Again, even Hopfield's much later survey of "Hopfield networks" (Scholarpedia, 2007)[AMH4] did not mention this.

2. Hinton vs Ivakhnenko and others

The related Boltzmann Machine paper by Ackley, Hinton, and Sejnowski (1985)[BM] was about learning internal representations in hidden units of neural networks (NNs).[S20] It didn't cite the first working algorithm for deep learning of internal representations by Ivakhnenko & Lapa (Ukraine, 1965).[DEEP1-2][HIN] It didn't cite Amari's separate work (1967-68)[GD1-2] on learning internal representations in deep NNs end-to-end through stochastic gradient descent (SGD). Not even the later surveys by the authors[S20][DL3][DLP] nor the *Scientific Background to the Nobel Prize in Physics 2024*[Nob24a] mention these origins of deep learning.

The Boltzmann machine-like Sherrington-Kirkpatrick model (1975)[SK75] is based on the general Edwards-Anderson model (1975)[EA75][EA21] (with random connections driving the network dynamics) and "learns" optimal weights J_{ij} for minimising *Free Energy*, where each point in the phase-diagram corresponds to an "internal representation." The 1985 Boltzmann Machine

paper[BM] fails to cite both papers. Even later surveys by Hinton failed to mention the 1975 Sherrington-Kirkpatrick model.

Hinton's Boltzmann Machine co-author Sejnowski[BM] was a student of Hopfield. He is also known for sending *"amicus curiae" ("friend of the court")* letters to award committees. He claims:[S20] *"In 1969, Minsky & Papert[M69] showed that shallow NNs without hidden layers are very limited and the field was abandoned until a new generation of neural network researchers took a fresh look at the problem in the 1980s."* This claim is echoed in the "Popular information" of the Nobel Committee: *"At the end of the 1960s, some discouraging theoretical results caused many researchers to suspect that these neural networks would never be of any real use."* Sejnowski also claimed:[S20b] *"Our goal was to try to take a network with multiple layers—an input layer, an output layer and layers in between—and make it learn. It was generally thought, because of early work that was done in AI in the 60s, that no one would ever find such a learning algorithm because it was just too mathematically difficult."*

However, this makes no sense, because the 1969 book[M69] addressed a "problem" of Gauss & Legendre's *shallow learning* with 1-layer NNs (~1800)[DL1-2][DLH] that had already been solved 4 years prior by Ivakhnenko & Lapa's popular *deep learning* method (1965),[DEEP1-2][DL2] and then also by Amari's stochastic gradient descent for multilayer perceptrons.[GD1-2] Minsky (who is cited by the *Scientific Background to the Nobel Prize in Physics 2024*[Nob24a]) was apparently unaware of this and failed to correct it later.[DLH][HIN][CONN21][DLP] Deep learning research was obviously alive and kicking in the 1960s-70s, especially outside of the Anglosphere.[DEEP1-2][GD1-3][CNN1][DL1-2][T22][DLP][DLH]

The Nobel Committee lauds Hinton et al.'s 2006 method for layer-wise pretraining of deep NNs.[UN4] However, this work neither cited the original layer-wise training of deep NNs by Ivakhnenko & Lapa (1965)[DEEP1-2] nor the original work on unsupervised pretraining of deep NNs (1991).[UN0-1][DLP]

Ivakhnenko's 1971 paper[DEEP2] already described a deep learning net with 8 layers.[DL2] This depth is comparable to the depth of Hinton's 2006 nets[UN4] published without comparison to the original work[DEEP1-2][DL2]—done when compute was millions of times more expensive. Given a training set of input vectors with corresponding target output vectors, layers are incrementally grown and trained by regression analysis. In a fine-tuning phase, superfluous hidden units are pruned with the help of a separate validation set.[DEEP2][DLH]

Hinton and Sejnowski have never cited the origins of deep learning in Ukraine and Japan in the 1960s and 1970s. None of the important algorithms for modern AI were invented by them.

3. Backpropagation

The Physics Nobel Committee[Nob24a] tries very hard to give the impression that modern pattern recognition and deep learning are somehow based on physics-inspired NNs, but they aren't. The well-known backpropagation technique (Linnainmaa, 1970)[BP1-5][BPA-C][DLP] is an efficient way of applying the chain rule (Leibniz, 1676)[LEI07-10][DLH] to big networks with differentiable nodes.[BP4] It is much more important for modern AI than physics-inspired equilibrium nets such as the so-

called "Hopfield network" and the "Boltzmann Machine" (which are irrelevant for modern AI applications mentioned by the Nobel Foundation[Nob24a]).

Backpropagation was also mentioned in the recent debate, although the Nobel Prize focuses on other things (otherwise the subsequent outcry would have been even greater). By 1985, compute was about 1,000 times cheaper than in 1970,[BP1][DLH] and the first desktop computers became accessible in wealthier academic labs. An experimental analysis of the known method[BP1-5] by Rumelhart et al. then demonstrated that backpropagation can yield useful internal representations in hidden layers of NNs.[RUM][DLH]

Hinton claimed that his co-author Rumelhart invented backpropagation.[AOI][HIN] The *Scientific Background to the Nobel Prize in Physics 2024,*[Nob24a] however, correctly cites Linnainmaa (1970),[BP1] who first published it, and Werbos (1982),[BP2] who first applied it to NNs.[DLP][DLH][DL1] It does *not* mention, however, that even later surveys by Hinton neither cited the original work by Linnainmaa[DLP] nor Amari's work (1967-68) on training networks with hidden layers through gradient descent.[GD1-2a] Reference [BP4] has a compact history of backpropagation and its precursors.

Note that as of 2024, Google Scholar (by Hinton's former employer) hallucinated 55k citations for a 1986 backpropagation paper by Rumelhart et al., simply adding 28k citations for the book in which it appeared.

(It is very remarkable that Amari and his student Saito[GD2][GD2a] applied stochastic gradient descent (SGD) to deep NNs in 1967-68 when compute was billions of times more expensive than today. However, SGD is not the same as backpropagation.)

4. Additional cases of plagiarism and misattribution

Many additional cases of plagiarism and incorrect attribution can be found in reference [DLP] which also contains most of the other references above. One can start with Sec. 3.

Two authors of a DeepMind team[DM4][NOB] each received 1/4 of the 2024 Nobel Prize in Chemistry for protein structure prediction through Alphafold[Nob24b] (disclaimer: DeepMind was co-founded by a student from my lab). They failed to cite important prior work by Baldi and Pollastri (2002):[DM4a] at a time when compute was roughly ten thousand times more expensive than in 2021, [DM4a] introduced a pipeline very similar to the one of Alphafold 2, using multiple sequence alignment (MSA) to predict the secondary protein structure with the help of a position-specific scoring matrix (PSSM) or a profile matrix, going beyond even earlier work of 1988.[DM4d][DM4e][Nob24b] The extra step (absent in Alphafold 2) was to predict the protein's topology, too. See also the important follow-up work of 2011.[DM4b] [DM4] also failed to cite Hochreiter et al.'s first successful application[HO07] of deep learning to protein folding (2007, using LSTM instead of MSA to construct a profile), as well as the essential prior work at TU Munich by Golkov et al (2016),[DM4c][DM4f] which had crucial aspects of AlphaFold: (1) identify homologous sequences in a database of proteins with known structure, (2) compute the co-evolution statistics using the homologous sequences, (3) train a graph NN to predict the protein contact map (that determines its 3D structure) directly from the co-evolution statistics, (4) demonstrate experimentally a significant boost in performance on the CASP dataset. Instead of the contact map, DeepMind (2021) predicted the distance map, and instead of graph CNNs, they used the

quadratic Transformer published in 2017[TR1] (the unnormalized linear Transformer had existed since 1991[ULTRA][FWP0-1,6][DLH]). DeepMind also used more training data and more compute for hyperparameter tuning etc.

Note that it is not acceptable to approve infringing work just because it has passed peer review once. Much past plagiarism has survived initial peer review before being detected, sometimes causing authors to lose their titles and jobs.

ACM is the organisation that hands out the Turing Awards. ACM's *Code of Ethics and Professional Conduct*[ACM18] states that computing professionals should *"credit the creators of ideas, inventions, work, and artifacts, and respect copyrights, patents, trade secrets, license agreements, and other methods of protecting authors' works."* Some of the awardees didn't.[DLP] The *"Policy for Honors Conferred by ACM"*[ACM23] mentions that ACM *"retains the right to revoke an Honor previously granted if ACM determines that it is in the best interests of the field to do so."* To my knowledge, the Nobel Committee has not published a similar code of ethics. The rules of scientific integrity, however, apply to all science awards.

In science, in the end, the facts must always win. As long as the facts have not yet won, it's not yet the end. No fancy award can ever change that.[HIN][T22][DLP][NOB] Nevertheless, in order to save the reputation of science prizes, awardees who have committed acts of plagiarism should be stripped of their awards.

The Romans already knew: magna est veritas et praevalebit (truth is mighty and will prevail).

Acknowledgments

Some of the material was taken from previous technical reports and AI Blog posts. [DLC][MIR][DEC] [LEI21][DAN] [DAN1][DL4][GPUCNN5,8][DLH] [MLP2][MOST] [UN][LSTMPG][BP4] [DL6a][HIN][T22][DLP] Thanks to many expert reviewers (including several famous neural net pioneers) for useful comments. Since science is about self-correction,[SV20] let me know under *juergen@idsia.ch* if you can spot any remaining error.

Partially annotated references

[ACM18] ACM Code of Ethics and Professional Conduct. Association for Computing Machinery (ACM), 2018. *Quote: "Computing professionals should therefore credit the creators of ideas, inventions, work, and artifacts, and respect copyrights, patents, trade secrets, license agreements, and other methods of protecting authors' works."*

[ACM23] Policy for Honors Conferred by ACM. Association for Computing Machinery (ACM), 2023. *Quote: "ACM also retains the right to revoke an Honor previously granted if ACM determines that it is in the best interests of the field to do so."* Copy in the Internet Archive (2023).

[AH1] Hentschel K. (1996) A. v. Brunn: Review of "100 Authors against Einstein" [March 13, 1931]. In: Hentschel K. (eds) Physics and National Socialism. Science Networks—Historical Studies, vol 18.

Birkhaeuser Basel. Link.

[AH2] F. H. van Eemeren, B. Garssen & B. Meuffels. The disguised abusive ad hominem empirically investigated: Strategic manoeuvring with direct personal attacks. Journal Thinking & Reasoning, Vol. 18, 2012, Issue 3, p. 344-364. Link.

[AH3] D. Walton (PhD Univ. Toronto, 1972), 1998. Ad hominem arguments. University of Alabama Press.

[AIB] J. Schmidhuber. AI Blog. Includes variants of chapters of the AI Book.

[AI51] Les Machines a Calculer et la Pensee Humaine: Paris, 8.-13. Januar 1951, Colloques internationaux du Centre National de la Recherche Scientifique; no. 37, Paris 1953. [*H. Bruderer rightly calls that the first conference on AI.*]

[AM24] S. I. Amari (2024). Personal communication.

[AMH0] S. I. Amari (1972). Characteristics of random nets of analog neuron-like elements. IEEE Trans. Syst. Man Cybernetics, 2, 643-657. *First published 1969 in Japanese, long before Wilson & Cowan's very similar work (1972-73).*

[AMH1] S. I. Amari (1972). Learning patterns and pattern sequences by self-organizing nets of threshold elements. IEEE Transactions, C 21, 1197-1206, 1972. PDF. *First publication of what was later sometimes called the Hopfield network[AMH2] or Amari-Hopfield Network,[AMH3] based on the (uncited) Lenz-Ising recurrent architecture.[L20][I25][T22][DLP] See also Little's work (1974-1980)[AMH1b-d] and this tweet.*

[AMH1b] W. A. Little. The existence of persistent states in the brain. Mathematical Biosciences, 19.1-2, p. 101-120, 1974. *Little uses Wannier's ideas of the 1940s[K41][W45] to express neural networks, and mentions the recurrent Ising model[L20][I25]on which the (uncited) Amari network[AMH1,2] is based.*

[AMH1c] W. A. Little and G. L. Shaw (1978). Analytic Study of the Memory Capacity of a Neural Network. Math Biosci. 39, 281–290 (1978). *This paper shows explicitly how to store-recall patterns with the Ising-Lenz model.*

[AMH1d] W. A. Little (1980). An Ising Model of a Neural Network. In: W. Jaeger, H. Rost, P. Tautu (eds), Biological Growth and Spread. Lecture Notes in Biomathematics, vol 38. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-61850-5_18

[AMH2] J. J. Hopfield (1982). Neural networks and physical systems with emergent collective computational abilities. Proc. of the National Academy of Sciences, vol. 79, pages 2554-2558, 1982. *The Hopfield network or Amari-Hopfield Network was first published in 1972 by Amari.[AMH1] [AMH2] did not cite [AMH1].*

[AMH3] A. P. Millan, J. J. Torres, J. Marro. How Memory Conforms to Brain Development. Front. Comput. Neuroscience, 2019

[AMH4] J. J. Hopfield (2007). Hopfield network. Scholarpedia, 2(5), 1977.

[AOI] M. Ford. Architects of Intelligence: The truth about AI from the people building it. Packt Publishing, 2018. (Preface to German edition by J. Schmidhuber.)

[ATT] J. Schmidhuber (AI Blog, 2020). 30-year anniversary of end-to-end differentiable sequential neural attention. Plus goal-conditional reinforcement learning. *Schmidhuber had both hard attention for foveas (1990) and soft attention in form of Transformers with linearized self-attention (1991-93).[FWP] Today, both types are very popular.*

[ATT0] J. Schmidhuber and R. Huber. Learning to generate focus trajectories for attentive vision. Technical Report FKI-128-90, Institut für Informatik, Technische Universität München, 1990. PDF.

[ATT1] J. Schmidhuber and R. Huber. Learning to generate artificial fovea trajectories for target detection. International Journal of Neural Systems, 2(1 & 2):135-141, 1991. Based on TR FKI-128-90, TUM, 1990. PDF. More.

[ATT2] J. Schmidhuber. Learning algorithms for networks with internal and external feedback. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, and G. E. Hinton, editors, *Proc. of the 1990 Connectionist Models Summer School*, pages 52-61. San Mateo, CA: Morgan Kaufmann, 1990. PS. (PDF.) *Reviewed by Dr. Hinton.*

[ATT3] H. Larochelle, G. E. Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine. NIPS 2010. *This work is very similar to [ATT0-2] which the authors did not cite. In fact, Hinton was the reviewer of a 1990 paper[ATT2] which summarised in its Section 5 Schmidhuber's early work on attention: the first implemented neural system for combining glimpses that jointly trains a recognition & prediction component with an attentional component (the fixation controller). Two decades later, Hinton wrote about his own work:[ATT3] "To our knowledge, this is the first implemented system for combining glimpses that jointly trains a recognition component ... with an attentional component (the fixation controller)." See [MIR](Sec. 9)[R4].*

[BM] D. Ackley, G. Hinton, T. Sejnowski (1985). A Learning Algorithm for Boltzmann Machines. Cognitive Science, 9(1):147-169. *This paper cited neither the Boltzmann machine-like Sherrington-Kirkpatrick model (1975)[SK75] based on the Edwards-Anderson model[EA75] nor Glauber[G63] nor the first working algorithms for deep learning of internal representations (Ivakhnenko & Lapa, 1965)[DEEP1-2][HIN] nor Amari's work (1967-68) [GD1-2] on learning internal representations in deep nets through stochastic gradient descent. Even later surveys by the authors[S20][DLC] failed to cite the prior art.[T22]*

[BPA] H. J. Kelley. Gradient Theory of Optimal Flight Paths. ARS Journal, Vol. 30, No. 10, pp. 947-954, 1960. *Precursor of modern backpropagation.[BP1-4]*

[BPB] A. E. Bryson. A gradient method for optimizing multi-stage allocation processes. Proc. Harvard Univ. Symposium on digital computers and their applications, 1961.

[BPC] S. E. Dreyfus. The numerical solution of variational problems. Journal of Mathematical Analysis and Applications, 5(1): 30-45, 1962.

[BP1] S. Linnainmaa. The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors. Master's Thesis (in Finnish), Univ. Helsinki, 1970. *See chapters 6-7 and FORTRAN code on pages 58-60.* PDF. See also BIT 16, 146-160, 1976. Link. *The first publication on "modern" backpropagation, also known as the reverse mode of automatic differentiation.*

[BP2] P. J. Werbos. Applications of advances in nonlinear sensitivity analysis. In R. Drenick, F. Kozin, (eds): System Modeling and Optimization: Proc. IFIP, Springer, 1982. PDF. *First application of backpropagation[BP1] to NNs (not yet in his 1974 thesis, as is sometimes claimed).*

[BP4] J. Schmidhuber (AI Blog, 2014; updated 2020). Who invented backpropagation? More.[DL2]

[BP5] A. Griewank (2012). Who invented the reverse mode of differentiation? Documenta Mathematica, Extra Volume ISMP (2012): 389-400.

[BPTT1] P. J. Werbos. Backpropagation through time: what it does and how to do it. Proceedings of the IEEE 78.10, 1550-1560, 1990.

[BPTT2] R. J. Williams and D. Zipser. Gradient-based learning algorithms for recurrent networks. In: Backpropagation: Theory, architectures, and applications, p 433, 1995.

[CNN1] K. Fukushima: Neural network model for a mechanism of pattern recognition unaffected by shift in position—Neocognitron. Trans. IECE, vol. J62-A, no. 10, pp. 658-665, 1979. *The first deep convolutional neural network architecture, with alternating convolutional layers and downsampling layers. In Japanese. English version: [CNN1+]. More in Scholarpedia.*

[CNN1+] K. Fukushima: Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological Cybernetics, vol. 36, no. 4, pp. 193-202 (April 1980). Link.

[CNN1a] A. Waibel. Phoneme Recognition Using Time-Delay Neural Networks. Meeting of IEICE, Tokyo, Japan, 1987. *First application of backpropagation[BP1][BP2] and weight-sharing to a 1-dimensional convolutional architecture.*

[CNN1a+] W. Zhang, J. Tanida, K. Itoh, Y. Ichioka. Shift-invariant pattern recognition neural network and its optical architecture. Proc. Annual Conference of the Japan Society of Applied Physics, 1988. *First "modern" backpropagation-trained 2-dimensional CNN.*

[CNN1b] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. J. Lang. Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 37, no. 3, pp. 328-339, March 1989. Based on [CNN1a].

[CNN1c] Bower Award Ceremony 2021: Jürgen Schmidhuber lauds Kunihiko Fukushima. YouTube video, 2021.

[CNN2] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel: Backpropagation Applied to Handwritten Zip Code Recognition, Neural Computation, 1(4):541-551, 1989.

[CNN3a] K. Yamaguchi, K. Sakamoto, A. Kenji, T. Akabane, Y. Fujimoto. A Neural Network for Speaker-Independent Isolated Word Recognition. First International Conference on Spoken Language Processing (ICSLP 90), Kobe, Japan, Nov 1990. *A 1-dimensional NN with convolutions using Max-Pooling instead of Fukushima's Spatial Averaging.[CNN1]*

[CNN3] Weng, J., Ahuja, N., and Huang, T. S. (1993). Learning recognition and segmentation of 3-D objects from 2-D images. Proc. 4th Intl. Conf. Computer Vision, Berlin, Germany, pp. 121-128. *A 2-dimensional CNN whose downsampling layers use Max-Pooling (which has become very popular) instead of Fukushima's Spatial Averaging.[CNN1]*

[CNN4] M. A. Ranzato, Y. LeCun: A Sparse and Locally Shift Invariant Feature Extractor Applied to Document Images. Proc. ICDAR, 2007

[CNN5a] S. Behnke. Learning iterative image reconstruction in the neural abstraction pyramid. International Journal of Computational Intelligence and Applications, 1(4):427-438, 1999.

[CNN5b] S. Behnke. Hierarchical Neural Networks for Image Interpretation, volume LNCS 2766 of Lecture Notes in Computer Science. Springer, 2003.

[CNN5c] D. Scherer, A. Mueller, S. Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In Proc. International Conference on Artificial Neural Networks (ICANN), pages 92-101, 2010.

[CONN21] Since November 2021: Comments on earlier versions of the report[T22] in the Connectionists Mailing List, perhaps the oldest mailing list on artificial neural networks. Link to the archive.

[CONN24] Since October 2024: messages to the Connectionists Mailing List, perhaps the oldest mailing list on artificial neural networks. Link to the archive.

[DAN] J. Schmidhuber (AI Blog, 2021). 10-year anniversary. In 2011, DanNet triggered the deep convolutional neural network (CNN) revolution. *Named after Schmidhuber's outstanding postdoc Dan Ciresan, it was the first deep and fast CNN to win international computer vision contests, and had a temporary monopoly on winning them, driven by a very fast implementation based on graphics processing units (GPUs). 1st superhuman result in 2011.[DAN1] Now everybody is using this approach.*

[DAN1] J. Schmidhuber (AI Blog, 2011; updated 2021 for 10th birthday of DanNet): First superhuman visual pattern recognition. *At the IJCNN 2011 computer vision competition in Silicon Valley, our artificial neural network called DanNet performed twice better than humans, three times better than the closest artificial competitor (by LeCun's team), and six times better than the best non-neural method.*

[DEC] J. Schmidhuber (AI Blog, 02/20/2020, updated 2025). The 2010s: Our Decade of Deep Learning / Outlook on the 2020s. *The recent decade's most important developments and industrial applications based on our AI, with an outlook on the 2020s, also addressing privacy and data markets.*

[DEEP1] Ivakhnenko, A. G. and Lapa, V. G. (1965). Cybernetic Predicting Devices. CCM Information Corporation. *First working Deep Learners with many layers, learning internal representations.*

[DEEP1a] Ivakhnenko, Alexey Grigorevich. The group method of data of handling; a rival of the method of stochastic approximation. Soviet Automatic Control 13 (1968): 43-55.

[DEEP2] Ivakhnenko, A. G. (1971). Polynomial theory of complex systems. IEEE Transactions on Systems, Man and Cybernetics, (4):364-378.

[DIST2] O. Vinyals, J. A. Dean, G. E. Hinton. Distilling the Knowledge in a Neural Network. Preprint arXiv:1503.02531 [stat.ML], 2015. *The authors did not cite Schmidhuber's original 1991 NN distillation*

procedure,[UN0-2][MIR](Sec. 2) *not even in the later patent application US20150356461A1. See also this tweet.*

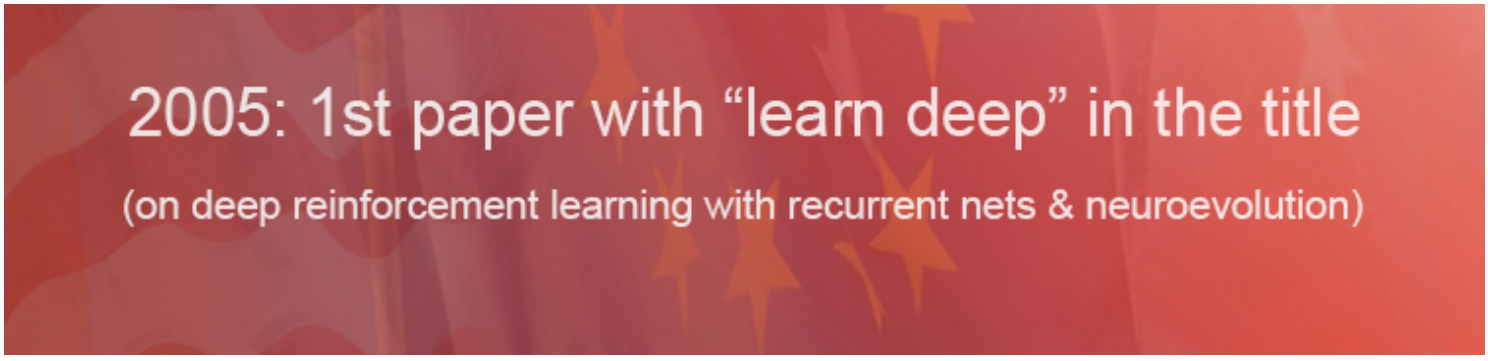
[DL1] J. Schmidhuber, 2015. Deep learning in neural networks: An overview. Neural Networks, 61, 85-117. More. *Got the first Best Paper Award ever issued by the journal Neural Networks, founded in 1988.*

[DL2] J. Schmidhuber, 2015. Deep Learning. Scholarpedia, 10(11):32832.

[DL3] Y. LeCun, Y. Bengio, G. Hinton (2015). Deep Learning. Nature 521, 436-444. HTML. *A "survey" of deep learning that does not mention the pioneering works of deep learning [T22].*

[DL3a] Y. Bengio, Y. LeCun, G. Hinton (2021). Turing Lecture: Deep Learning for AI. Communications of the ACM, July 2021. HTML. Local copy (HTML only). *Another "survey" of deep learning that does not mention the pioneering works of deep learning [T22].*

[DL4] J. Schmidhuber (AI Blog, 2017). Our impact on the world's most valuable public companies: Apple, Google, Microsoft, Facebook, Amazon... *By 2015-17, neural nets developed in Schmidhuber's labs 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. Examples: greatly improved (CTC-based) speech recognition on all Android phones, greatly improved machine translation through Google Translate and Facebook (over 4 billion LSTM-based translations per day), Apple's Siri and Quicktype on all iPhones, the answers of Amazon's Alexa, etc. Google's 2019 on-device speech recognition (on the phone, not the server) is still based on LSTM.*



[DL6] F. Gomez and J. Schmidhuber. Co-evolving recurrent neurons learn deep memory POMDPs. In *Proc. GECCO'05*, Washington, D. C., pp. 1795-1802, ACM Press, New York, NY, USA, 2005. PDF.

[DL6a] J. Schmidhuber (AI Blog, Nov 2020). 15-year anniversary: 1st paper with "learn deep" in the title (2005). Our deep reinforcement learning & neuroevolution solved problems of depth 1000 and more.[DL6]

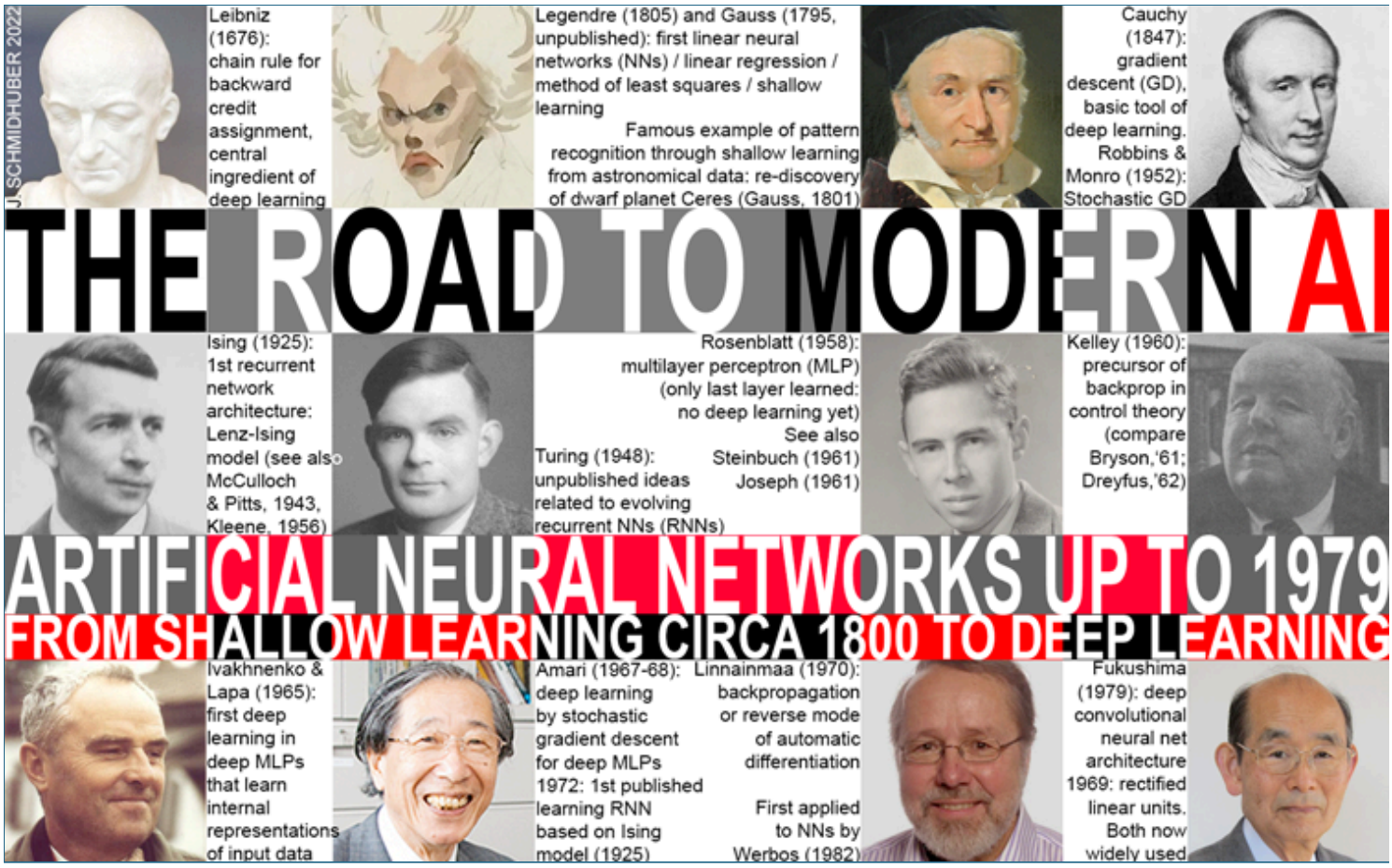
Soon after its publication, everybody started talking about "deep learning." Causality or correlation?

[DL7] "Deep Learning ... moving beyond shallow machine learning since 2006!" Web site deeplearning.net of Y. Bengio's MILA (2015, retrieved May 2020; compare the version in the Internet Archive), referring to Hinton's[UN4] and Bengio's[UN5] *unsupervised* pre-training for deep NNs[UN4] (2006) although this type of deep learning dates back to Schmidhuber's work of 1991.[UN1-2][UN] Compare Sec. II & XVII & III.

[DLC] J. Schmidhuber (AI Blog, June 2015). Critique of Paper by self-proclaimed[DLC1-2] "Deep Learning Conspiracy" (Nature 521 p 436). *The inventor of an important method should get credit for inventing it. She may not always be the one who popularizes it. Then the popularizer should get credit for popularizing it (but not for inventing it).*

[DLC1] Y. LeCun. IEEE Spectrum Interview by L. Gomes, Feb 2015. *Quote: "A lot of us involved in the resurgence of Deep Learning in the mid-2000s, including Geoff Hinton, Yoshua Bengio, and myself—the so-called 'Deep Learning conspiracy' ..."*

[DLC2] M. Bergen, K. Wagner (2015). Welcome to the AI Conspiracy: The 'Canadian Mafia' Behind Tech's Latest Craze. Vox recode, 15 July 2015. *Quote: "... referred to themselves as the 'deep learning conspiracy.' Others called them the 'Canadian Mafia.'"*



[DLH] J. Schmidhuber (AI Blog, 2022). Annotated History of Modern AI and Deep Learning. Technical Report IDSIA-22-22, IDSIA, Lugano, Switzerland, 2022. Preprint arXiv:2212.11279. Tweet of 2022.

[DLP] J. Schmidhuber (AI Blog, 2023). How 3 Turing awardees republished key methods and ideas whose creators they failed to credit. Technical Report IDSIA-23-23, Swiss AI Lab IDSIA, 14 Dec 2023. The piece is aimed at people who are not aware of the numerous AI priority disputes, but are willing to check the facts (see tweet).

[DM4] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli & D. Hassabis. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583-589, 2021.

[DM4a] P. Baldi, G. Pollastri. A machine learning strategy for protein analysis. IEEE Intelligent Systems 17.2 (2002): 28-35.

[DM4b] P. Di Lena, K. Nagata, and P. Baldi. Deep Architectures for Protein Contact Map Prediction. Bioinformatics, 28, 2449-2457, (2012).

[DM4c] V. Golkov, M. J. Skwark, A. Golkov, A. Dosovitskiy, T. Brox, J. Meiler, D. Cremers (2016). Protein contact prediction from amino acid co-evolution using convolutional networks for graph-valued images. Advances in Neural Information Processing Systems (NeurIPS), Barcelona, 2016.

[DM4d] N. Qian and T.J. Sejnowski (1988). Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. 1988, 202, 865-884.

[DM4e] H. Bohr, J. Bohr, S. Brunak, R.M.J. Cotterill, B. Lautrup, L. Norskov, O.H. Olsen, S.B. Petersen (1988). Protein secondary structure and homology by neural networks. The α-helices in rhodopsin. FEBS Lett. 1988, 241, 223-228.

[DM4f] D. Cremers (July 2025). LinkedIn post on the Nobel Prize for AlphaFold.

[Drop1] S. J. Hanson (1990). A Stochastic Version of the Delta Rule, PHYSICA D,42, 265-272. *What's now called "dropout" is a variation of the stochastic delta rule—compare preprint arXiv:1808.03578, 2018. Note also that Ivakhnenko and Lapa (1965-1970)[DEEP1-2][DLH] already pruned hidden units from their deep networks.*

[Drop2] N. Frazier-Logue, S. J. Hanson (2020). The Stochastic Delta Rule: Faster and More Accurate Deep Learning Through Adaptive Weight Noise. Neural Computation 32(5):1018-1032.

[Drop3] J. Hertz, A. Krogh, R. Palmer (1991). Introduction to the Theory of Neural Computation. Redwood City, California: Addison-Wesley Pub. Co., pp. 45-46.

[Drop4] N. Frazier-Logue, S. J. Hanson (2018). Dropout is a special case of the stochastic delta rule: faster and more accurate deep learning. Preprint arXiv:1808.03578, 2018.

[Drop5] S. A. Janowsky (1989). Pruning versus clipping in neural networks. Phys. Rev. A 39, 6600, 1989. https://doi.org/10.1103/PhysRevA.39.6600

[EA75] S. F. Edwards and P. W. Anderson (1975). Theory of Spin Glasses. Journal of Physics F: Metal Physics, 5, 965, https://doi.org/10.1088/0305-4608/5/5/017. *The Boltzmann machine-like Sherrington-Kirkpatrick model (1975) [SK75] is based on the Edward-Anderson model.*

[EA21] G. Parisi (2021). Nobel Lecture. *Slide with Edward-Anderson Hamiltonian[EA75] at 6:12.*

[FAKE] H. Hopf, A. Krief, G. Mehta, S. A. Matlin. Fake science and the knowledge crisis: ignorance can be fatal. Royal Society Open Science, May 2019. *Quote: "Scientists must be willing to speak out when they see false information being presented in social media, traditional print or broadcast press" and "must speak out against false information and fake science in circulation and forcefully contradict public figures who promote it."*

[FAKE2] L. Stenflo. Intelligent plagiarists are the most dangerous. Nature, vol. 427, p. 777 (Feb 2004). *Quote: "What is worse, in my opinion, ..., are cases where scientists rewrite previous findings in different words, purposely hiding the sources of their ideas, and then during subsequent years forcefully claim that they have discovered new phenomena."*

[FWP] J. Schmidhuber (AI Blog, 26 March 2021, updated 2025). 26 March 1991: Neural nets learn to program neural nets with fast weights—like Transformer variants. 2021: New stuff! *30-year anniversary of a now popular alternative[FWP0-1] to recurrent NNs. A slow feedforward NN learns by gradient descent to program the changes of the fast weights[FAST,FASTa] of another NN, separating memory and control like in traditional computers. Such Fast Weight Programmers[FWP0-6,FWPMETA1-8] can learn to memorize past data, e.g., by computing fast weight changes through additive outer products of self-invented activation patterns[FWP0-1] (now often called keys and values for self-attention[TR1-6]). The similar Transformers[TR1-2] combine this with projections and softmax and are now widely used in natural language processing. For long input sequences, their efficiency was improved through Transformers with linearized self-attention[TR5-6] which are formally equivalent to Schmidhuber's 1991 outer product-based Fast Weight Programmers (apart from normalization), now called unnormalized linear Transformers.[ULTRA] In 1993, he introduced the attention terminology[FWP2] now used in this context,[ATT] and extended the approach to RNNs that program themselves. See tweet of 2022.*

[FWP0] J. Schmidhuber. Learning to control fast-weight memories: An alternative to recurrent nets. Technical Report FKI-147-91, Institut für Informatik, Technische Universität München, 26 March 1991. PDF. *First paper on fast weight programmers that separate storage and control: a slow net learns by gradient descent to compute weight changes of a fast net. The outer product-based version (Eq. 5) is now known as an unnormalized linear Transformer or "Transformer with linearized self-attention."[FWP]*

[FWP1] J. Schmidhuber. Learning to control fast-weight memories: An alternative to recurrent nets. Neural Computation, 4(1):131-139, 1992. Based on [FWP0]. PDF. HTML. Pictures (German). See tweet of 2022 for 30-year anniversary.

[FWP2] J. Schmidhuber. Reducing the ratio between learning complexity and number of time-varying variables in fully recurrent nets. In Proceedings of the International Conference on Artificial Neural Networks, Amsterdam, pages 460-463. Springer, 1993. PDF. *First recurrent NN-based fast weight*

programmer using outer products (a recurrent extension of the 1991 unnormalized linear Transformer), introducing the terminology of learning "internal spotlights of attention."

[FWP6] I. Schlag, K. Irie, J. Schmidhuber. Linear Transformers Are Secretly Fast Weight Programmers. ICML 2021. Preprint: arXiv:2102.11174.

[G63] R. J Glauber (1963). Time-dependent statistics of the Ising model. Journal of Mathematical Physics, 4(2):294-307, 1963.

[GD'] C. Lemarechal. Cauchy and the Gradient Method. Doc Math Extra, pp. 251-254, 2012.

[GD''] J. Hadamard. Memoire sur le probleme d'analyse relatif a Vequilibre des plaques elastiques encastrees. Memoires presentes par divers savants estrangers à l'Academie des Sciences de l'Institut de France, 33, 1908.

[GDa] Y. Z. Tsypkin (1966). Adaptation, training and self-organization automatic control systems, Avtomatika I Telemekhanika, 27, 23-61. *On gradient descent-based on-line learning for non-linear systems.*

[GDb] Y. Z. Tsypkin (1971). Adaptation and Learning in Automatic Systems, Academic Press, 1971. *On gradient descent-based on-line learning for non-linear systems.*

[GD1] S. I. Amari (1967). A theory of adaptive pattern classifier, IEEE Trans, EC-16, 279-307 (Japanese version published in 1965). PDF. *Probably the first paper on using stochastic gradient descent[STO51-52] for learning in multilayer neural networks (without specifying the specific gradient descent method now known as reverse mode of automatic differentiation or backpropagation[BP1]).*

[GD2] S. I. Amari (1968). Information Theory—Geometric Theory of Information, Kyoritsu Publ., 1968 (in Japanese). OCR-based PDF scan of pages 94-135 (see pages 119-120). *Contains computer simulation results for a five layer network (with 2 modifiable layers) which learns internal representations to classify non-linearily separable pattern classes.* See also this tweet.

[GD2a] H. Saito (1967). Master's thesis, Graduate School of Engineering, Kyushu University, Japan. *Implementation of Amari's 1967 stochastic gradient descent method for multilayer perceptrons.[GD1] (S.*

Amari, personal communication, 2021.)

[GD3] S. I. Amari (1977). Neural Theory of Association and Concept Formation. Biological Cybernetics, vol. 26, p. 175-185, 1977. *See Section 3.1 on using gradient descent for learning in multilayer networks.*

[GUE96] J. Schmidhuber, S. Hochreiter (1996). Guessing can outperform many long time lag algorithms. Technical Note IDSIA-19-96, IDSIA, May 1996.

[GPUNN] Oh, K.-S. and Jung, K. (2004). GPU implementation of neural networks. Pattern Recognition, 37(6):1311-1314. *Speeding up traditional NNs on GPU by a factor of 20.*

[GPUCNN] K. Chellapilla, S. Puri, P. Simard. High performance convolutional neural networks for document processing. International Workshop on Frontiers in Handwriting Recognition, 2006. *Speeding up shallow CNNs on GPU by a factor of 4.*

[GPUCNN1] D. C. Ciresan, U. Meier, J. Masci, L. M. Gambardella, J. Schmidhuber. Flexible, High Performance Convolutional Neural Networks for Image Classification. *International Joint Conference on Artificial Intelligence (IJCAI-2011, Barcelona)*, 2011. PDF. ArXiv preprint. *Speeding up deep CNNs on GPU by a factor of 60. Used to win four important computer vision competitions 2011-2012 before others won any with similar approaches.*

[GPUCNN2] D. C. Ciresan, U. Meier, J. Masci, J. Schmidhuber. A Committee of Neural Networks for Traffic Sign Classification. *International Joint Conference on Neural Networks (IJCNN-2011, San Francisco)*, 2011. PDF. HTML overview. *First superhuman performance in a computer vision contest, with half the error rate of humans, and one third the error rate of the closest competitor.[DAN1] This led to massive interest from industry.*

[GPUCNN3] D. C. Ciresan, U. Meier, J. Schmidhuber. Multi-column Deep Neural Networks for Image Classification. Proc. *IEEE Conf. on Computer Vision and Pattern Recognition CVPR 2012*, p 3642-3649, July 2012. PDF. Longer TR of Feb 2012: arXiv:1202.2745v1 [cs.CV]. More.

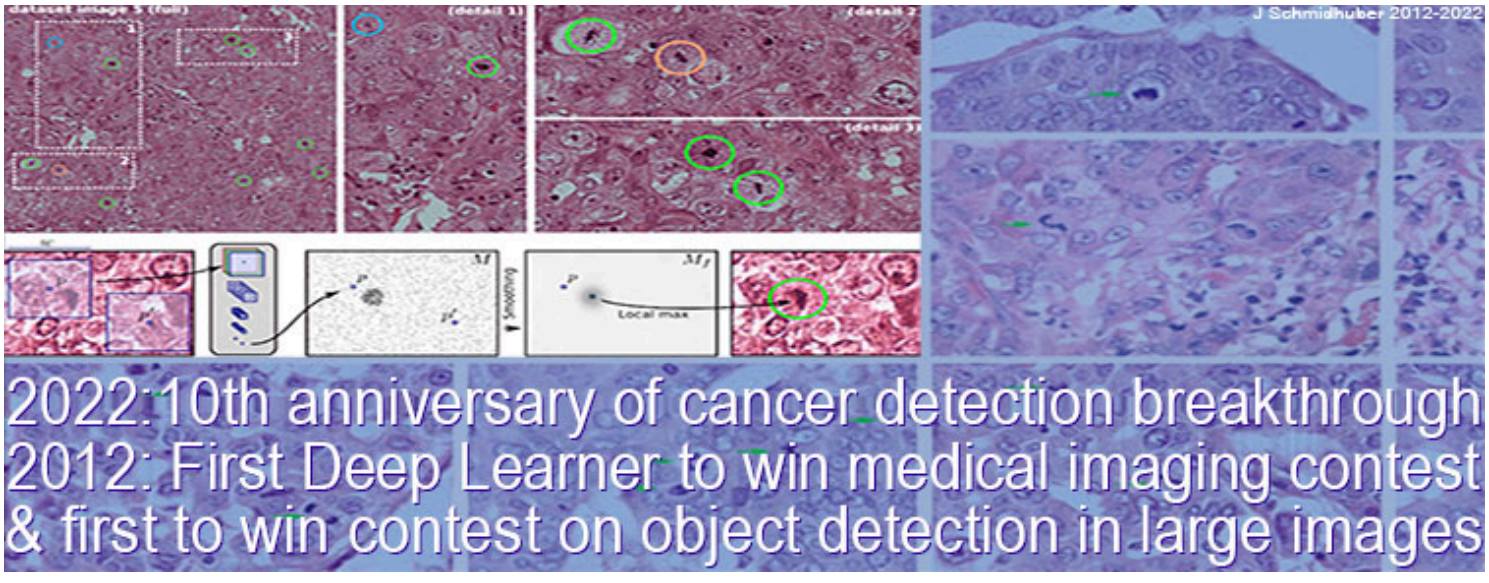
[GPUCNN4] A. Krizhevsky, I. Sutskever, G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. NIPS 25, MIT Press, Dec 2012. PDF. *This paper describes AlexNet, which is similar to the earlier DanNet,[DAN,DAN1][R6] the first pure deep CNN to win computer vision contests in 2011[GPUCNN2-3,5] (AlexNet and VGG Net[GPUCNN9] followed in 2012-2014). [GPUCNN4] emphasizes benefits of Fukushima's ReLUs (1969)[RELU1] and dropout (a variant of Hanson 1990 stochastic delta rule)[Drop1-4] but neither cites the original work[RELU1][Drop1] nor the basic CNN architecture (Fukushima, 1979).[CNN1]*

[GPUCNN5] J. Schmidhuber (AI Blog, 2017; updated 2021 for 10th birthday of DanNet): History of computer vision contests won by deep CNNs since 2011. DanNet won 4 of them in a row before the similar AlexNet/VGG Net and the Resnet (a Highway Net with open gates) joined the party. Today, deep CNNs are standard in computer vision.

[GPUCNN6] J. Schmidhuber, D. Ciresan, U. Meier, J. Masci, A. Graves. On Fast Deep Nets for AGI Vision. In Proc. Fourth Conference on Artificial General Intelligence (AGI-11), Google, Mountain View, California, 2011. PDF.

[GPUCNN7] D. C. Ciresan, A. Giusti, L. M. Gambardella, J. Schmidhuber. Mitosis Detection in Breast Cancer Histology Images using Deep Neural Networks. MICCAI 2013. PDF.

[GPUCNN8] J. Schmidhuber (AI Blog, 2017; updated 2021 for 10th birthday of DanNet). First deep learner to win a contest on object detection in large images— first deep learner to win a medical imaging contest (2012). Link. *How the Swiss AI Lab IDSIA used GPU-based CNNs to win the ICPR 2012 Contest on Mitosis Detection and the MICCAI 2013 Grand Challenge.*



[GPUCNN9] K. Simonyan, A. Zisserman. Very deep convolutional networks for large-scale image recognition. Preprint arXiv:1409.1556 (2014).

[GRO78] S. Grossberg (1978). Competition, decision, and consensus. Journal of Mathematical Analysis and Applications, 66, 470-493. PDF.

[GRO83] M.A. Cohen, S. Grossberg (1983). Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, 815-826. *According to Grossberg [CONN24], this work was actually first submitted to the Journal of Cybernetics in 1980, just when that journal stopped publishing. It built on earlier theorems [GRO78] proving global limits for the "Additive Model" which introduced a Lyapunov functional to help prove convergence.*

[GRO20] Grossberg, S. (2020). A path towards Explainable AI and autonomous adaptive intelligence: Deep Learning, Adaptive Resonance, and models of perception, emotion, and action. Frontiers in Neurobotics, June 25, 2020. HTML.

[GRO21] Grossberg, S. (2021). Conscious mind, resonant brain: how each brain makes a mind. Oxford University Press.

[HOP84] J. J. Hopfield (1984). Neurons with graded response have collective computational properties like those of two-state neurons." Proceedings of the national academy of sciences 81.10 (1984): 3088-3092.

[H86] J. L. van Hemmen (1986). Spin-glass models of a neural network. Phys. Rev. A 34, 3435, 1 Oct 1986.

[H88] H. Sompolinsky (1988). Statistical Mechanics of Neural Networks. Physics Today 41, 12, 70, 1988.

[HIN] J. Schmidhuber (AI Blog, 2020). Critique of Honda Prize for Dr. Hinton. *Science must not allow corporate PR to distort the academic record. See also this tweet.*

[HO07] S. Hochreiter, M. Heusel, K. Obermayer. Fast model-based protein homology detection without alignment. Bioinformatics 23(14):1728-36, 2007. *Successful application of deep learning to protein folding problems, through an LSTM that was orders of magnitude faster than competing methods.*

[HYB12] Hinton, G. E., Deng, L., Yu, D., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Process. Mag., 29(6):82-97. *This work did not cite the earlier LSTM[LSTM0-6] trained by Connectionist Temporal Classification (CTC, 2006).[CTC] CTC-LSTM was successfully applied to speech in 2007[LSTM4] (also with hierarchical LSTM stacks[LSTM14]) and became the first superior end-to-end neural speech recogniser that outperformed the state of the art, dramatically improving Google's speech recognition.[GSR][GSR15][DL4] This was very different from previous hybrid methods since the late 1980s which combined NNs and traditional approaches such as hidden Markov models (HMMs).[BW][BRI][BOU] [HYB12] still used the old hybrid approach and did not compare it to CTC-LSTM. Later, however, Hinton switched to LSTM, too.[LSTM8]*

[I24] E. Ising (1925). Beitrag zur Theorie des Ferro- und Paramagnetismus. Dissertation, 1924.

[I25] E. Ising (1925). Beitrag zur Theorie des Ferromagnetismus. Z. Phys., 31 (1): 253-258, 1925. *The first non-learning recurrent NN architecture (the Ising model or Lenz-Ising model) was introduced and analyzed by physicists Ernst Ising and Wilhelm Lenz in the 1920s.*[L20][I25][K41][W45][T22] *It settles into an equilibrium state in response to input conditions, and is the foundation of the first published learning RNNs.*[AMH1-2]

[JC67] J. Cowan. Statistical Mechanics of Neural Networks. AD0658886, Chicago University. Link.

[K41] H. A. Kramers and G. H. Wannier (1941). Statistics of the Two-Dimensional Ferromagnet. Phys. Rev. 60, 252 and 263, 1941.

[L20] W. Lenz (1920). Beitrag zum Verständnis der magnetischen Erscheinungen in festen Körpern. Physikalische Zeitschrift, 21:613-615. See also [I25].

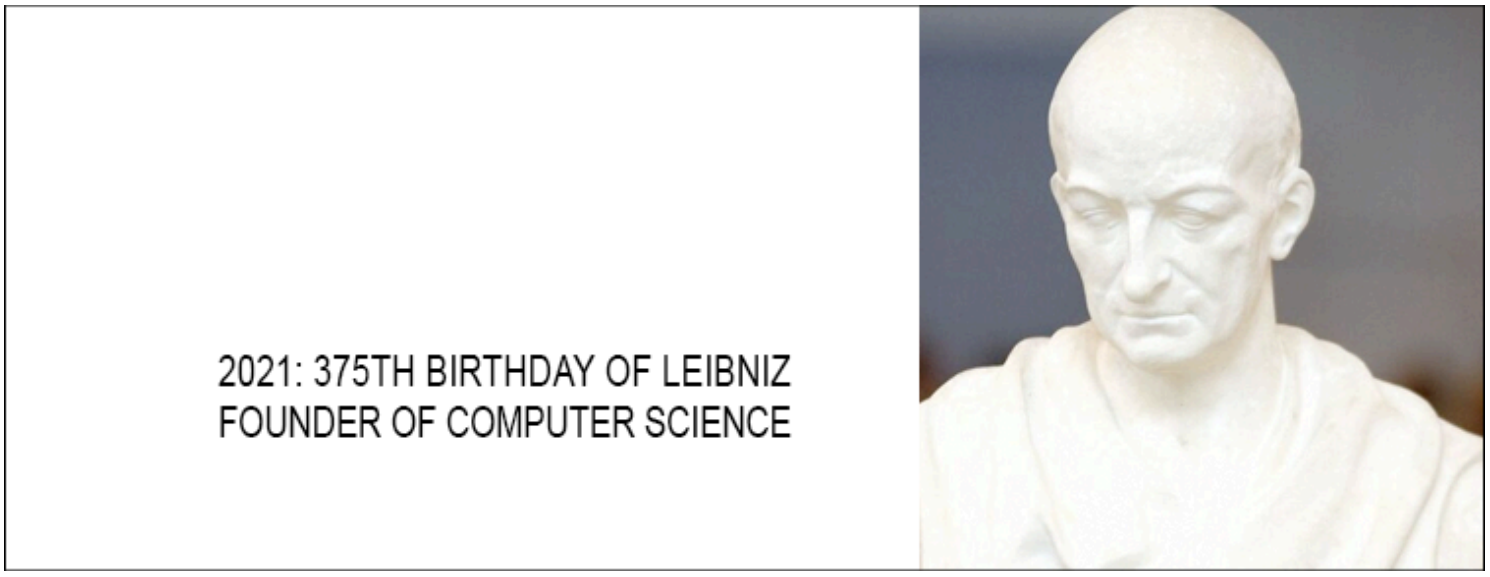
[LAN] J. L. Ba, J. R.Kiros, G. E. Hinton. Layer Normalization. arXiv:1607.06450, 2016.

[LEI07] J. M. Child (translator), G. W. Leibniz (Author). The Early Mathematical Manuscripts of Leibniz. Merchant Books, 2007. *See p. 126: the chain rule appeared in a 1676 memoir by Leibniz.*

[LEI10] O. H. Rodriguez, J. M. Lopez Fernandez (2010). A semiotic reflection on the didactics of the Chain rule. The Mathematics Enthusiast: Vol. 7 : No. 2 , Article 10. DOI: https://doi.org/10.54870/1551-3440.1191.

[LEI21] J. Schmidhuber (AI Blog, 2021). 375th birthday of Leibniz, founder of computer science.

[LEI21a] J. Schmidhuber (2021). Der erste Informatiker. Wie Gottfried Wilhelm Leibniz den Computer erdachte. (The first computer scientist. How Gottfried Wilhelm Leibniz conceived the computer.) Frankfurter Allgemeine Zeitung (FAZ), 17/5/2021. FAZ online: 19/5/2021.



[LIT21] M. L. Littman (2021). Collusion Rings Threaten the Integrity of Computer Science Research. Communications of the ACM, Vol. 64 No. 6, p. 43-44, June 2021.

[LSTM0] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. TR FKI-207-95, TUM, August 1995. PDF.

[LSTM1] S. Hochreiter, J. Schmidhuber. Long Short-Term Memory. Neural Computation, 9(8):1735-1780, 1997. PDF. Based on [VAN1][LSTM0][GUE96]. More.

[LSTM2] F. A. Gers, J. Schmidhuber, F. Cummins. Learning to Forget: Continual Prediction with LSTM. Neural Computation, 12(10):2451-2471, 2000. PDF. *The "vanilla LSTM architecture" with forget gates that*

everybody is using today, e.g., in Google's Tensorflow.

[LSTM3] A. Graves, J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18:5-6, pp. 602-610, 2005. PDF.

[LSTM4] S. Fernandez, A. Graves, J. Schmidhuber. An application of recurrent neural networks to discriminative keyword spotting. *Intl. Conf. on Artificial Neural Networks ICANN'07,* 2007. PDF.

[LSTM5] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, J. Schmidhuber. A Novel Connectionist System for Improved Unconstrained Handwriting Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 5, 2009. PDF.

[LSTM6] A. Graves, J. Schmidhuber. Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. NIPS'22, p 545-552, Vancouver, MIT Press, 2009. PDF.

[LSTM7] J. Bayer, D. Wierstra, J. Togelius, J. Schmidhuber. Evolving memory cell structures for sequence learning. Proc. ICANN-09, Cyprus, 2009. PDF.

[LSTM8] A. Graves, A. Mohamed, G. E. Hinton. Speech Recognition with Deep Recurrent Neural Networks. ICASSP 2013, Vancouver, 2013. PDF. *Based on [LSTM1-2,4,14][CTC].*

[LSTM9] O. Vinyals, L. Kaiser, T. Koo, S. Petrov, I. Sutskever, G. Hinton. Grammar as a Foreign Language. Preprint arXiv:1412.7449 [cs.CL].

[LSTM10] A. Graves, D. Eck and N. Beringer, J. Schmidhuber. Biologically Plausible Speech Recognition with LSTM Neural Nets. In J. Ijspeert (Ed.), First Intl. Workshop on Biologically Inspired Approaches to Advanced Information Technology, Bio-ADIT 2004, Lausanne, Switzerland, p. 175-184, 2004. PDF.

[LSTM11] N. Beringer and A. Graves and F. Schiel and J. Schmidhuber. Classifying unprompted speech by retraining LSTM Nets. In W. Duch et al. (Eds.): Proc. Intl. Conf. on Artificial Neural Networks ICANN'05, LNCS 3696, pp. 575-581, Springer-Verlag Berlin Heidelberg, 2005.

[LSTM12] D. Wierstra, F. Gomez, J. Schmidhuber. Modeling systems with internal state using Evolino. In Proc. of the 2005 conference on genetic and evolutionary computation (GECCO), Washington, D. C., pp. 1795-1802, ACM Press, New York, NY, USA, 2005. Got a GECCO best paper award.

[LSTM13] F. A. Gers and J. Schmidhuber. LSTM Recurrent Networks Learn Simple Context Free and Context Sensitive Languages. IEEE Transactions on Neural Networks 12(6):1333-1340, 2001. PDF.

[LSTM14] S. Fernandez, A. Graves, J. Schmidhuber. Sequence labelling in structured domains with hierarchical recurrent neural networks. In Proc. IJCAI 07, p. 774-779, Hyderabad, India, 2007 (talk). PDF.

[LSTM15] A. Graves, J. Schmidhuber. Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks. *Advances in Neural Information Processing Systems 22, NIPS'22,* p 545-552, Vancouver, MIT Press, 2009. PDF.

[LSTM16] M. Stollenga, W. Byeon, M. Liwicki, J. Schmidhuber. Parallel Multi-Dimensional LSTM, With Application to Fast Biomedical Volumetric Image Segmentation. Advances in Neural Information Processing Systems (NIPS), 2015. Preprint: arxiv:1506.07452.

[LSTM17] J. A. Perez-Ortiz, F. A. Gers, D. Eck, J. Schmidhuber. Kalman filters improve LSTM network performance in problems unsolvable by traditional recurrent nets. Neural Networks 16(2):241-250, 2003. PDF.

[LSTMPG] J. Schmidhuber (AI Blog, Dec 2020). 10-year anniversary of our journal paper on deep reinforcement learning with policy gradients for LSTM (2007-2010). *Recent famous applications:*

DeepMind's Starcraft player (2019) and OpenAI's dextrous robot hand & Dota player (2018)—Bill Gates called this a huge milestone in advancing AI.

[LSTM-RL] B. Bakker, F. Linaker, J. Schmidhuber. Reinforcement Learning in Partially Observable Mobile Robot Domains Using Unsupervised Event Extraction. In Proceedings of the 2002 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2002), Lausanne, 2002. PDF.

[LSTMGRU] J. Chung, C. Gulcehre, K. Cho, Y. Bengio (2014). Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. Preprint arXiv:1412.3555 [cs.NE]. *The so-called gated recurrent units (GRU) are actually a variant of the vanilla LSTM architecture[LSTM2] (2000) which the authors did not cite although this work[LSTM2] was the one that introduced gated recurrent units. They cited only the 1997 LSTM[LSTM1] which did not yet have "forget gates."[LSTM2] Furthermore, Schmidhuber's team automatically evolved lots of additional LSTM variants and topologies already in 2009[LSTM7] without changing the name of the basic method. (Margin note: GRU cells lack an important gate and can neither learn to count[LSTMGRU2] nor learn simple non-regular languages;[LSTMGRU2] they also do not work as well for challenging translation tasks, according to Google Brain.[LSTMGRU3])*

[LSTMGRU2] G. Weiss, Y. Goldberg, E. Yahav. On the Practical Computational Power of Finite Precision RNNs for Language Recognition. Preprint arXiv:1805.04908.

[LSTMGRU3] D. Britz et al. (2017). Massive Exploration of Neural Machine Translation Architectures. Preprint arXiv:1703.03906

[M69] M. Minsky, S. Papert. Perceptrons (MIT Press, Cambridge, MA, 1969). *A misleading "history of deep learning" goes more or less like this: "In 1969, Minsky & Papert[M69] showed that shallow NNs without hidden layers are very limited and the field was abandoned until a new generation of neural network researchers took a fresh look at the problem in the 1980s."[S20] However, the 1969 book[M69] addressed a "problem" of Gauss & Legendre's shallow learning with 1-layer NNs (~1800)[DL1-2] that had already been solved 4 years prior by Ivakhnenko & Lapa's popular deep learning method,[DEEP1-2][DL2] and then also by Amari's SGD for MLPs.[GD1-2] Minsky was apparently unaware of this and failed to correct it later.[HIN](Sec. I)[T22](Sec. XIII)[DLP]*

[MIR] J. Schmidhuber (AI Blog, Oct 2019, updated 2025). Deep Learning: Our Miraculous Year 1990-1991. Preprint arXiv:2005.05744. *The deep learning neural networks (NNs) of our team have revolutionised pattern recognition & machine learning & AI. Many of the basic ideas behind this revolution were published within fewer than 12 months in our "Annus Mirabilis" 1990-1991 at TU Munich, including principles of (1) LSTM, the most cited AI of the 20th century (based on constant error flow through residual connections); (2) ResNet, the most cited AI of the 21st century (based on our LSTM-inspired Highway Network, 10 times deeper than previous NNs); (3) GAN (for artificial curiosity and creativity); (4) Transformer (the T in*

ChatGPT—see the 1991 Unnormalized Linear Transformer); (5) Pre-training for deep NNs (the P in ChatGPT); (6) NN distillation (see DeepSeek); (7) recurrent World Models, and more.

[MLP1] D. C. Ciresan, U. Meier, L. M. Gambardella, J. Schmidhuber. Deep Big Simple Neural Nets For Handwritten Digit Recognition. Neural Computation 22(12): 3207-3220, 2010. ArXiv Preprint. *Showed that plain backprop for deep standard NNs is sufficient to break benchmark records, without any unsupervised pre-training.*

[MLP2] J. Schmidhuber (AI Blog, Sep 2020). 10-year anniversary of supervised deep learning breakthrough (2010). No unsupervised pre-training. *By 2010, when compute was 100 times more expensive than today, both the feedforward NNs[MLP1] and the earlier recurrent NNs of Schmidhuber's team were able to beat all competing algorithms on important problems of that time.*

[MLP3] J. Schmidhuber (AI Blog, 2025). 2010: Breakthrough of end-to-end deep learning (no layer-by-layer training, no unsupervised pre-training). The rest is history. *By 2010, when compute was 1000 times more expensive than in 2025, both our feedforward NNs[MLP1] and our earlier recurrent NNs were able to beat all competing algorithms on important problems of that time. This deep learning revolution quickly spread from Europe to North America and Asia.*

[MOST] J. Schmidhuber (AI Blog, 2021, updated 2025). The most cited neural networks all build on work done in my labs: *1. Long Short-Term Memory (LSTM), the most cited AI of the 20th century. 2. ResNet (open-gated Highway Net), the most cited AI of the 21st century. 3. AlexNet & VGG Net (the similar but earlier DanNet of 2011 won 4 image recognition challenges before them). 4. GAN (an instance of Adversarial Artificial Curiosity of 1990). 5. Transformer variants—see the 1991 unnormalised linear Transformer (ULTRA). Foundations of Generative AI were published in 1991: the principles of GANs (now used for deepfakes), Transformers (the T in ChatGPT), Pre-training for deep NNs (the P in ChatGPT), NN distillation, and the famous DeepSeek—see the tweet.*

[NAK72] K. Nakano. Associatron—A Model of Associative Memory. IEEE Transactions on Systems, Man, and Cybernetics, SMC-2:3 p. 380-388, 1972.

[NASC1] J. Schmidhuber. First Pow(d)ered flight / plane truth. Correspondence, *Nature, 421 p 689, Feb 2003.*

[NASC2] J. Schmidhuber. Zooming in on aviation history. Correspondence, *Nature, vol 566, p 39, 7 Feb 2019*.

[NASC3] J. Schmidhuber. The last inventor of the telephone. Letter, *Science, 319, no. 5871, p. 1759, March 2008.*

[NASC4] J. Schmidhuber. Turing: Keep his work in perspective. Correspondence, *Nature, vol 483, p 541, March 2012*, doi:10.1038/483541b.

[NASC5] J. Schmidhuber. Turing in Context. Letter, *Science, vol 336, p 1639, June 2012*. (On Gödel, Zuse, Turing.) See also comment on response by A. Hodges (DOI:10.1126/science.336.6089.1639-a)

[NASC6] J. Schmidhuber. Colossus was the first electronic digital computer. Correspondence, *Nature, 441 p 25, May 2006.*

[NASC7] J. Schmidhuber. Turing's impact. Correspondence, *Nature, 429 p 501, June 2004*

[NASC8] J. Schmidhuber. Prototype resilient, self-modeling robots. Correspondence, *Science, 316, no. 5825 p 688, May 2007.*

[NASC9] J. Schmidhuber. Comparing the legacies of Gauss, Pasteur, Darwin. Correspondence, *Nature, vol 452, p 530, April 2008.*

[NAT1] J. Schmidhuber. Citation bubble about to burst? Nature, vol. 469, p. 34, 6 January 2011. HTML.

[NHE] J. Schmidhuber. The Neural Heat Exchanger. Oral presentations since 1990 at various universities including TUM and the University of Colorado at Boulder. Also in In S. Amari, L. Xu, L. Chan, I. King, K. Leung, eds., Proceedings of the Intl. Conference on Neural Information Processing (1996), pages 194-197, Springer, Hongkong. Link. *Proposal of a biologically more plausible deep learning algorithm that—unlike backpropagation—is local in space and time. Inspired by the physical heat exchanger: inputs "heat up" while being transformed through many successive layers, targets enter from the other end of the deep pipeline and "cool down."*

[NOB] J. Schmidhuber. A Nobel Prize for Plagiarism. Technical Report IDSIA-24-24 (7 Dec 2024, updated 2025). *Sadly, the Nobel Prize in Physics 2024 for Hopfield & Hinton is a Nobel Prize for plagiarism. They republished methodologies for artificial neural networks developed in Ukraine and Japan by Ivakhnenko and Amari in the 1960s & 1970s, as well as other techniques, without citing the original papers. Even in later surveys, they didn't credit the original inventors (thus turning what may have been unintentional plagiarism into a deliberate form). None of the important algorithms for modern Artificial Intelligence were created by Hopfield & Hinton. See also popular tweet1, tweet2, and LinkedIn post.*

[Nob10] J. Schmidhuber (2010). Evolution of National Nobel Prize Shares in the 20th Century. Technical Report, IDSIA & USI & SUPSI, Switzerland, 14 September 2010. Preprint arXiv:1009.2634v1. (Compare ScienceNews Blog, 1 Oct 2010.)

[Nob24a] The Nobel Committee for Physics (2024). Scientific Background to the Nobel Prize in Physics 2024. PDF. (Local copy.)

[Nob24b] The Nobel Committee for Chemistry (2024). Scientific Background to the Nobel Prize in Chemistry 2024. PDF.

[NOBieee] M. S. Smith (2024). Why the Nobel Prize in Physics Went to AI Research. IEEE Spectrum, 2024.

[NOBnat] E. Gibney & D. Castelvecchi (2024). Physics Nobel scooped by machine-learning pioneers. Nature, 8 Oct 2024.

[NYT1] NY Times article by J. Markoff, Nov. 27, 2016: When A.I. Matures, It May Call Jürgen Schmidhuber 'Dad'

[PLA1] E. Creamer. Authors file a lawsuit against OpenAI for unlawfully ‘ingesting’ their books. The Guardian, 5 Jul 2023.

[PLAG1] Oxford's guide to types of plagiarism (2021). *Quote: "Plagiarism may be intentional or reckless, or unintentional."* Copy in the Internet Archive. Local copy.

[PLAG2] Jackson State Community College (2022). Unintentional Plagiarism. Copy in the Internet Archive.

[PLAG3] R. L. Foster. Avoiding Unintentional Plagiarism. Journal for Specialists in Pediatric Nursing; Hoboken Vol. 12, Iss. 1, 2007.

[PLAG4] N. Das. Intentional or unintentional, it is never alright to plagiarize: A note on how Indian universities are advised to handle plagiarism. Perspect Clin Res 9:56-7, 2018.

[PLAG5] InfoSci-OnDemand (2023). What is Unintentional Plagiarism? Copy in the Internet Archive.

[PLAG6] Copyrighted.com (2022). How to Avoid Accidental and Unintentional Plagiarism (2023). Copy in the Internet Archive. *Quote: "May it be accidental or intentional, plagiarism is still plagiarism."*

[PLAG7] Cornell Review, 2024. Harvard president resigns in plagiarism scandal. 6 January 2024.

Relevant threads with many comments at reddit.com/r/MachineLearning, the largest machine learning forum with over 800k subscribers in 2019 (note that my name is often misspelled):

[R1] Reddit/ML, 2019. Hinton, LeCun, Bengio receive ACM Turing Award. *This announcement contains more comments about Schmidhuber than about any of the awardees.*

[R2] Reddit/ML, 2019. J. Schmidhuber really had GANs in 1990.

[R3] Reddit/ML, 2019. NeurIPS 2019 Bengio Schmidhuber Meta-Learning Fiasco. *Schmidhuber started metalearning (learning to learn—now a hot topic) in 1987[META1][META] long before Bengio who suggested in public at N(eur)IPS 2019 that he did it before Schmidhuber.*

[R4] Reddit/ML, 2019. Five major deep learning papers by G. Hinton did not cite similar earlier work by J. Schmidhuber.

[R5] Reddit/ML, 2019. The 1997 LSTM paper by Hochreiter & Schmidhuber has become the most cited deep learning research paper of the 20th century.

[R6] Reddit/ML, 2019. DanNet, the CUDA CNN of Dan Ciresan in J. Schmidhuber's team, won 4 image recognition challenges prior to AlexNet.

[R7] Reddit/ML, 2019. J. Schmidhuber on Seppo Linnainmaa, inventor of backpropagation in 1970.

[R8] Reddit/ML, 2019. J. Schmidhuber on Alexey Ivakhnenko, godfather of deep learning 1965.

[R58] Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6):386. *This paper not only described single layer perceptrons, but also deeper multilayer perceptrons (MLPs). Although these MLPs did not yet have deep learning, because only the last layer learned,[DL1][DLH] Rosenblatt basically had what much later was rebranded as Extreme Learning Machines (ELMs) without proper attribution.[ELM1-2][CONN21][T22]*

[R61] Joseph, R. D. (1961). Contributions to perceptron theory. PhD thesis, Cornell Univ.

[R62] Rosenblatt, F. (1962). Principles of Neurodynamics. Spartan, New York.

[RELU1] K. Fukushima (1969). Visual feature extraction by a multilayered network of analog threshold elements. IEEE Transactions on Systems Science and Cybernetics. 5 (4): 322-333. doi:10.1109/TSSC.1969.300225. *This work introduced rectified linear units or ReLUs.*

[RELU2] C. v. d. Malsburg (1973). Self-Organization of Orientation Sensitive Cells in the Striate Cortex. Kybernetik, 14:85-100, 1973. *See Table 1 for rectified linear units or ReLUs. Possibly this was also the first work on applying an EM algorithm to neural nets.*

[RUM] DE Rumelhart, GE Hinton, RJ Williams (1985). Learning Internal Representations by Error Propagation. TR No. ICS-8506, California Univ San Diego La Jolla Inst for Cognitive Science. Later version published as: Learning representations by back-propagating errors. Nature, 323, p. 533-536 (1986). *This experimental analysis of backpropagation did not cite the origin of the method,[BP1-5] also known as the reverse mode of automatic differentiation. The paper also failed to cite the first working algorithms for deep learning of internal representations (Ivakhnenko & Lapa, 1965)[DEEP1-2][HIN][DLH] as well as Amari's work (1967-68)[GD1-2] on learning internal representations in deep nets through stochastic gradient descent. Even later surveys by the authors[DL3,3a] failed to cite the prior art.[T22][DLP][NOB]*

[S93] D. Sherrington (1993). Neural networks: the spin glass approach. North-Holland Mathematical Library, vol 51, 1993, p. 261-291.

[S20] T. Sejnowski. The unreasonable effectiveness of deep learning in artificial intelligence. PNAS, January 28, 2020. Link. *A misleading "history of deep learning" which goes more or less like this: "In 1969, Minsky & Papert[M69] showed that shallow NNs without hidden layers are very limited and the field was abandoned until a new generation of neural network researchers took a fresh look at the problem in the 1980s."[S20] However, the 1969 book[M69] addressed a "problem" of Gauss & Legendre's shallow learning with 1-layer NNs (~1800)[DL1-2][DLH] that had already been solved 4 years prior by Ivakhnenko & Lapa's popular deep learning method,[DEEP1-2][DL2] and then also by Amari's SGD for MLPs.[GD1-2] Minsky was apparently unaware of this and failed to correct it later.[HIN](Sec. I)[T22](Sec. XIII)[DLP] Deep learning research was alive and kicking in the 1960s-70s, especially outside of the Anglosphere.[DEEP1-2][GD1-3][CNN1][DL1-2][T22][DLH]*

[S20b] C. S. Smith. Do Machines Dream? An Interview With Terry Sejnowski on Boltzmann Machines and The Brain. Paperspace, Digital Ocean, 2020. Link.

[S80] B. Speelpenning (1980). Compiling Fast Partial Derivatives of Functions Given by Algorithms. PhD thesis, Department of Computer Science, University of Illinois, Urbana-Champaign.

[STO51] H. Robbins, S. Monro (1951). A Stochastic Approximation Method. The Annals of Mathematical Statistics. 22(3):400, 1951.

[STO52] J. Kiefer, J. Wolfowitz (1952). Stochastic Estimation of the Maximum of a Regression Function. The Annals of Mathematical Statistics. 23(3):462, 1952.

[SK75] D. Sherrington, S. Kirkpatrick (1975). Solvable Model of a Spin-Glass. Phys. Rev. Lett. 35, 1792, 1975. *See also [EA75].*

[ST61] K. Steinbuch. Die Lernmatrix. (The learning matrix.) Kybernetik, 1(1):36-45, 1961.

[ST95] W. Hilberg (1995). Karl Steinbuch, ein zu Unrecht vergessener Pionier der künstlichen neuronalen Systeme. (Karl Steinbuch, an unjustly forgotten pioneer of artificial neural systems.) Frequenz, 49(1995)1-2.

[SV20] S. Vazire (2020). A toast to the error detectors. Let 2020 be the year in which we value those who ensure that science is self-correcting. Nature, vol 577, p 9, 2/2/2020.

[T19] ACM's justification of the 2018 A.M. Turing Award (announced in 2019). WWW link. Local copy 1 (HTML only). Local copy 2 (HTML only). [T22] debunks this justification.

[T20a] J. Schmidhuber (AI Blog, 25 June 2020). Critique of 2018 Turing Award for Drs. Bengio & Hinton & LeCun. A precursor of [T22].

[T21v1] J. Schmidhuber. Scientific Integrity, the 2021 Turing Lecture, and the 2018 Turing Award for Deep Learning. Technical Report IDSIA-77-21 (v1), IDSIA, 24 Sep 2021.

[T22] J. Schmidhuber (AI Blog, 2022). Scientific Integrity and the History of Deep Learning: The 2021 Turing Lecture, and the 2018 Turing Award. Technical Report IDSIA-77-21, IDSIA, Lugano, Switzerland, 2022. *Debunking [T19] and [DL3a] .*

[TR1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin (2017). Attention is all you need. NIPS 2017, pp. 5998-6008. *This paper introduced the name "Transformers" for a now widely used NN type. It did not cite the 1991 publication on what's now called unnormalized linear Transformers with "linearized self-attention."[ULTRA] Schmidhuber also introduced the now popular attention terminology in 1993.[ATT][FWP2][R4] See tweet of 2022 for 30-year anniversary.*

[TR2] J. Devlin, M. W. Chang, K. Lee, K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint arXiv:1810.04805.

[TR5] A. Katharopoulos, A. Vyas, N. Pappas, F. Fleuret. Transformers are RNNs: Fast autoregressive Transformers with linear attention. In Proc. Int. Conf. on Machine Learning (ICML), July 2020.

[TR6] K. Choromanski, V. Likhosherstov, D. Dohan, X. Song, A. Gane, T. Sarlos, P. Hawkins, J. Davis, A. Mohiuddin, L. Kaiser, et al. Rethinking attention with Performers. In Int. Conf. on Learning Representations (ICLR), 2021.

[TUR] A. M. Turing. On computable numbers, with an application to the Entscheidungsproblem. Proceedings of the London Mathematical Society, Series 2, 41:230-267. Received 28 May 1936. Errata appeared in Series 2, 43, pp 544-546 (1937). *2nd explicit proof that the Entscheidungsproblem (decision problem) does not have a general solution.*

[TUR1] A. M. Turing. Intelligent Machinery. Unpublished Technical Report, 1948. Link. In: Ince DC, editor. Collected works of AM Turing - Mechanical Intelligence. Elsevier Science Publishers, 1992.

[TUR2] A. M. Turing (1952). The Chemical Basis of Morphogenesis. Philosophical Transactions of the Royal Society of London 237 (641):37-72.

[TUR21] J. Schmidhuber (AI Blog, Sep 2021). Turing Oversold. It's not Turing's fault, though.

[ULTRA] References on the 1991 unnormalized linear Transformer (ULTRA): original tech report (1991) [FWP0]. Journal publication (1992) [FWP1]. Recurrent ULTRA extension (1993) introducing the terminology of learning "internal spotlights of attention" [FWP2]. Modern *"quadratic"* Transformer (2017: *"attention is all you need"*) scaling *quadratically* in input size [TR1]. Papers of 2020-21 using the terminology *"linearized attention"* for more efficient *"linear Transformers"* that scale *linearly* [TR5,TR6]. 2021 paper [FWP6] pointing out that ULTRA dates back to 1991 [FWP0] when compute was a million times more expensive. ULTRA overview (2021) [FWP]. See the T in ChatGPT! See also surveys [DLH] [DLP], 2022 tweet for ULTRA's 30-year anniversary, and 2024 tweet.

[UN] J. Schmidhuber (AI Blog, 2021). 30-year anniversary. 1991: First very deep learning with unsupervised pre-training. First neural network distillation. *Unsupervised hierarchical predictive coding (with self-supervised target generation) finds compact internal representations of sequential data to facilitate downstream deep learning. The hierarchy can be distilled into a single deep neural network (suggesting a simple model of conscious and subconscious information processing). 1993: solving problems of depth >1000.*

[UN0] J. Schmidhuber. Neural sequence chunkers. Technical Report FKI-148-91, Institut für Informatik, Technische Universität München, April 1991. PDF. *Unsupervised/self-supervised learning and predictive coding is used in a deep hierarchy of recurrent neural networks (RNNs) to find compact internal representations of long sequences of data, across multiple time scales and levels of abstraction. Each RNN tries to solve the pretext task of predicting its next input, sending only unexpected inputs to the next RNN above. The resulting compressed sequence representations greatly facilitate downstream supervised deep learning such as sequence classification. By 1993, the approach solved problems of depth 1000 [UN2] (requiring 1000 subsequent computational stages/layers—the more such stages, the deeper the learning). A variant collapses the hierarchy into a single deep net. It uses a so-called conscious chunker RNN which attends to unexpected events that surprise a lower-level so-called subconscious automatiser RNN. The chunker learns to understand the surprising events by predicting them. The automatiser uses a neural knowledge distillation procedure to compress and absorb the formerly conscious insights and*

behaviours of the chunker, thus making them subconscious. The systems of 1991 allowed for much deeper learning than previous methods. More.

[UN1] J. Schmidhuber. Learning complex, extended sequences using the principle of history compression. Neural Computation, 4(2):234-242, 1992. Based on TR FKI-148-91, TUM, 1991.[UN0] PDF. *First working Deep Learner based on a deep RNN hierarchy (with different self-organising time scales), overcoming the vanishing gradient problem through unsupervised pre-training and predictive coding (with self-supervised target generation). Also: compressing or distilling a teacher net (the chunker) into a student net (the automatizer) that does not forget its old skills—such approaches are now widely used. See also this tweet. More.*

[UN2] J. Schmidhuber. Habilitation thesis, TUM, 1993. PDF. *An ancient experiment on "Very Deep Learning" with credit assignment across 1200 time steps or virtual layers and unsupervised / self-supervised pre-training for a stack of recurrent NN can be found here (depth > 1000).*

[UN4] G. E. Hinton, R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, Vol. 313. no. 5786, pp. 504—507, 2006. PDF. *This work describes unsupervised layer-wise pre-training of stacks of feedforward NNs (FNNs) called Deep Belief Networks (DBNs). However, this work neither cited the original layer-wise training of deep NNs by Ivakhnenko & Lapa (1965)[DEEP1-2] nor the 1991 unsupervised pre-training of stacks of more general recurrent NNs (RNNs)[UN0-2] which introduced the first NNs shown to solve very deep problems. The 2006 justification of the authors was essentially the one Schmidhuber used for the 1991 RNN stack: each higher level tries to reduce the description length (or negative log probability) of the data representation in the level below.[HIN][T22][MIR] This can greatly facilitate very deep downstream learning.[UN0-2]*

[UN5] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle. Greedy layer-wise training of deep networks. Proc. NIPS 06, pages 153-160, Dec. 2006. *The comment under reference[UN4] applies here as well.*

[VAN1] S. Hochreiter. Untersuchungen zu dynamischen neuronalen Netzen. Diploma thesis, TUM, 1991 (advisor J. Schmidhuber). PDF. *More on the Fundamental Deep Learning Problem.*

[VAN2] Y. Bengio, P. Simard, P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE TNN 5(2), p 157-166, 1994. *Results are essentially identical to those of Schmidhuber's diploma student Hochreiter (1991).[VAN1] Even after a common publication,[VAN3] the first author of [VAN2] published papers[VAN4-5] that cited only their own [VAN2] but not the original work.*

[VAN3] S. Hochreiter, Y. Bengio, P. Frasconi, J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer and J. F. Kolen, eds., A Field Guide to Dynamical Recurrent Neural Networks. IEEE press, 2001. PDF.

[VAN4] Y. Bengio. Neural net language models. Scholarpedia, 3(1):3881, 2008. Link.

[VAN5] R. Pascanu, T. Mikolov, Y. Bengio. On the difficulty of training Recurrent Neural Networks. ICML 2013.

[W45] G. H. Wannier (1945). The Statistical Problem in Cooperative Phenomena. Rev. Mod. Phys. 17, 50.

[WER87] P. J. Werbos. Building and understanding adaptive systems: A statistical/numerical approach to factory automation and brain research. IEEE Transactions on Systems, Man, and Cybernetics, 17, 1987.

[WER89] P. J. Werbos. Backpropagation and neurocontrol: A review and prospectus. In IEEE/INNS International Joint Conference on Neural Networks, Washington, D.C., volume 1, pages 209-216, 1989.