Book Review

No Harm Intended

Marvin L. Minsky and Seymour A. Papert.
Perceptrons: An Introduction to Computational Geometry, Expanded Edition. Cambridge, MA: MIT Press, 1988. 292pp.

Reviewed by Jordan B. Pollack

The Authors are professors at the Massachusetts Institute of Technology, Minsky in Electrical Engineering and Papert in Applied Mathematics. Minsky's major research interests are artificial intelligence (AI) and theories of computation, and Papert's are cybernetics and child development

The Reviewer received his Ph.D. in Computer Science from the University of Illinois in 1987, and is currently an Assistant Professor of Computer and Information Science at the Ohio State University. His research involves the use of connectionist models in traditional AI tasks.

Introduction

By 1969, the date of the publication of Perceptrons, AI was not operating in an ivory-tower vacuum. Money was at stake. And while this pressured the field into a preference for short-term achievement, it also put a premium on claims that the sponsor's investment would bear fruits beyond the immediate product. It [was] almost inevitable for AI to appropriate our work as proof that neural networks were universally bad. We did not think of our work as killing Snow White... (Papert, 1988 p. 7)

Research on Connectionist and Neural Networks has recently awakened, and is now expecting a valuable kiss from Prince DARPA.[1] In fact, 1988 may be the year of connectionism's loss of innocence. Not only has 1988 seen the publication of articles, including Papert's, about the symbolic/connectionist debate in *Daedelus*, but also the publication of a provocative special anti-connectionist issue of *Cognition*. Finally, to top it off, MIT Press has re-issued *Perceptrons*, with a new prologue and epilogue, titled, respectively, ''The View from 1988,'' and ''The New Connectionism''

Because this is the third edition of the book, I will focus attention on the new material,[2] which is more political than scientific. First, however, for those who have not read *Perceptrons*, I will summarize its scientific contents. This work, in retrospect, will most likely be viewed as the greatest contribution to the cognitive sciences made by Minsky or Papert, so it is still important to understand their findings. The book is indeed valuable to those who intend to work on neural networks from the viewpoint of

---

[1] The Defense Advanced Research Projects Agency has recently announced a major funding initiative in neural networks.

[2] The new edition refers to reviews of first edition by Newell (1969) and Block (1970).

theoretical computer science, although the field of computational complexity of parallel algorithms has evolved quite rapidly, and independently, of Minsky and Papert's work.

The book uses some modern mathematics, such as topology and group theory, to explain the computational and adaptive capacities and limitations of simple networks of threshold logic units, which perform pattern-recognition tests by thresholding linear combinations of binary input vectors with numerical weights.

First, Minsky and Papert defined the *order* of a predicate as the size of the largest conjunction in the minimal sum-of-products logical form for that predicate. Thus while both conjunction and alternation are predicates of order 1, exclusive-or is a predicate of order 2. The generalization of exclusive-or to more than 2 inputs is parity, which is not of finite order: a sum of products to represent parity of n inputs has at least one term of size n. As a predicate of non-finite order is scaled, then, there is no limit to the fan-in of linear threshold units.

Using arguments of symmetry, Minsky and Papert then showed, with their *Group Invariance Theorem*, that linear threshold functions which are invariant under a permutation group can be transformed into a function whose coefficients depend only on the group; a major result is that the only linear (i.e., order 1) functions invariant under such transitive groups as scaling, translation and rotation are simple size or area measures. Attempts to use linear threshold functions for, say, optical character recognition under these transitive conditions are thus doomed to failure.

After a cataloguing of the orders of various geometric functions, Minsky and Papert focused on the problems of learning. They showed that as various predicates scale, the sizes of coefficients can grow exponentially, thus leading to systems of impractical memory requirements needing unbounded cycles of a convergence procedure.

No one, in fact, disputes the contribution made by *Perceptrons* to the study of networks of neuron-like units, or disputes the authors' results on how such networks scale. The reviewer considers, as deep truth, their lesson that successful learning is more dependent on what is to be learned than on details of the learning mechanism. It is therefore unsurprising that very hard cases can be devised for supposedly general purpose learning mechanisms (Blum & Rivest, 1988; Judd, 1987; Kolen, 1988), up to the problem of categorizing whether or not inputs are ''halting tapes for Turing Machines.'' (p. 226.)

The two major issues in dispute, between Minsky and Papert and the modern PDP movement, are first, a matter of responsibility, and second, the interpretation of recent results in the field.

Malice Aforethought?

Minsky and Papert surrounded their 1969 mathematical tract with fairly negative judgements and loaded terms, such as the following quotes, which have been used as evidence (Dreyfus & Dreyfus, 1988; Rumelhart & Zipser, 1986) that they actually intended to stifle research on perceptron-like models.

> *Perceptrons have been widely publicized as ‘‘pattern recognition’’ or ‘‘learning’’ machines and as such have been discussed in a large number of books, journal articles, and voluminous ‘‘reports.’’ Most of this writing ... is without scientific value. (p. 4)*

> *We do not see that any good can come of experiments which pay no attention to limiting factors that will assert themselves as soon as the small model is scaled up to a usable size. (p. 18)*

> *[We] became involved with a somewhat therapeutic compulsion: to dispel what we feared to be the first shadows of a ‘‘holistic’’ or ‘‘Gestalt’’ misconception that would threaten to haunt the fields of engineering and artificial intelligence... (p. 20)*

> *There is no reason to suppose that any of these virtues carry over to the many layered version. Nevertheless, we consider it to be an important research problem to elucidate (or reject) our intuitive judgement that the extension is sterile. (p. 231)*

Despite these pronouncements, in 1988, Minsky and Papert wish to deny their responsibility, or, at least, their intentionality, in bringing about the decade-long connectionist winter:

> *One popular version is that the publication of our book so discouraged research on learning in network machines that a promising line of research was interrupted. Our version is that progress had already come to a virtual halt because of the lack of adequate basic theories. (p. xii)*

Their explanation is that two problems, recognized by Samuel (1967), and a fundamental principle, stopped all research on machine learning, perceptrons included. The first problem is the apportionment of credit (or blame) for incorrect global behavior of a system to specific elements in an interdependent system of such elements. The second problem is the creation of novel and useful higher-order abstractions. The basic law is that a system must be able to potentially represent what it wants to learn, and Minsky and Papert assert that perceptrons "had no way to represent the knowledge required for solving certain problems." (p. xiii).

The credit assignment and novel feature problems are, indeed, central to the ‘‘new connectionism.’’ For example, by relaxing the activation function from a threshold to a sigmoidal, Rumelhart, Hinton, and Williams (1986) were able to demonstrate credit assignment in multiple layered networks, and Hinton (1986) showed how such networks, when appropriately constrained, could develop semantically interpretable higher-order elements.

And, although the old new connectionism, circa 1980 - 1985 (cf. Waltz and Pollack (1985)), was fairly weak on representational issues, there is currently quite a focus on representation in the new new connectionism (Pollack, 1988; Touretzky, 1986; Touretzky & Geva, 1987) One reason that it has taken connectionism several years to rediscover the law of representational adequacy is that there was no body of work in AI involving issues of knowledge representation in neural models.

But the real problem which terminated the research viability of perceptron-like models was the problem of scaling. Minsky and Papert asserted that as such learning models based on gradient descent in weight space were scaled up, they would be impractical due to local minima, extremely large weights and a concurrent growth in convergence time.

So, were they responsible for killing Snow White? No, since intention and action are separable, they were no more responsible than Bill, who, intending to kill his uncle, is ‘‘so nervous and excited [when driving] that he accidentally runs over and kills a pedestrian, who happens to be his uncle’’ (Searle 1980, p. 51). If Minsky and Papert did not intend to stifle the field of neural networks, then, perhaps, they would act in accordance with their new motto: ‘‘We see no reason to choose sides.’’ (p. xiv).

Malice Afterthought

Unfortunately, the new epilogue is not in accordance with this relaxed position. Instead, it is an attempt to make good their assertion, in the prologue, that ‘‘little of significance [has] changed since 1969.’’ They take on the new connectionism, as embodied in the Parallel Distributed Processing (PDP) volumes (McClelland et al, 1986; Rumelhart et al, 1986), although they focus mainly on Chapter 8 (Rumelhart et al., 1986b).

Some of the new writing echoes their earlier battle rhetoric. In a long paragraph, for example, they discuss the new connectionism as a ‘‘new and heated ... intellectually aggressive movement to establish a paradigm.’’ It’s not just a movement, it’s a ‘‘connectionist revolution’’, fomented by the professional agitators Rumelhart, McClelland and their ‘‘14 collaborators,’’ who published a ‘‘manifesto.’’. Even though they only criticize Chapter 8, they hope that the reader will ‘‘seize the general’’ point. Finally, the authors wrap themselves in the ‘‘flag’’ of unsolvable problems (pp. 247-248).

Minsky and Papert point out, rightly, that Rumelhart et al have no result, like the Perceptron Convergence Theorem (Rosenblatt, 1962), that the generalized delta rule is guaranteed to find solutions when they exist, and that ‘‘little has been proved about the range of problems upon which [the generalized delta rule] works both efficiently and dependably.’’ (p. 260.)

Since, despite the fact that they "consider[ed] it an important research problem," they have no new results about multi-layered perceptrons in the 19 years since the book's first publication, they, in fact, know just how hard it is to answer questions like this. But they also miss the fact that such questions are not of central concern to psychologists, such as Rumelhart and McClelland, who might even use first-order systems in pursuit of a cognitive model (Rumelhart & McClelland, 1986) which is empirically testable and deniable (Pinker & Prince, 1988).

Minsky and Papert go on to give a cursory reexamination of scaling issues for the generalized delta rule, and, in fine form, conclude that:

> *We do not know if this fact is significant, but many of the small examples described in PDP could have been solved as quickly by means of exhaustive search -- that is by systematically assigning and testing all combinations of small integer weights. (p. 262.)*

We do know, however, that generalizations from single pieces of evidence (i.e. PDP Chapter 8), are usually poor. While one could indeed solve the credit assignment problem for the seven weights of the exclusive-or network using exhaustive search, one could not do it for a text-to-phoneme network (Sejnowski & Rosenberg, 1987) with its 10^5 weights or an autonomous vehicle controller with 10^6 (Pomerleau, 1988).

On the other hand, we also know that for certain classes of very hard problems, there are no algorithms which scale significantly better than exhaustive search, unless they are approximations. But there are many such impractical and approximate methods which are in practical use worldwide. Despite the fact that the Traveling Salesman Problem is NP-complete, salespersons do travel, and often optimally. Many problems which come up in operations research and automated manufacturing, such as integer programming, are similarly combinatorial in nature, but this is not used as evidence to stop work on these problems.

Large numbers and long convergence times may be the cost of doing business in machine learning, until, and this is really doubtful, somebody discovers the secret of Alchemical Intelligence, turning the lead of NP into the gold of P.

Fairy Tales and Double Standards

But AI has not discovered the secret P=NP key. And it has remained free to work on toy problems, in the United States at least, without ever being held hostage to the double standard of scaling up.

For example, work continues on AI systems which rely on schema-based memory, ignoring the combinatorics of the schema selection problem. The notion of a theorem prover carrying out logical inference on thousands of axioms and millions of propositions is hardly viable as a cognitive architecture. No rule-based system has survived more than a few thousand rules, and the problems of maintaining such large rule-bases are unsolved. And qualitative reasoning in the domain of naive physics, which seems to have dominated much of the mainstream publication of AI for the past few years, ignores the issue of computational tractability altogether.[3] Thus, scaling is a disease which affects all kinds of AI, including the symbolic kind that Minsky and Papert were trying to protect from the holistic or gestalt misconception.

Scaling is, in fact, a very important aspect of current research in connectionism. For example, Rosenfeld and Touretzky (1987) analyzed the scaling behavior of coarse coded distributed representations, Mjolsness, Sharp, and Alpert (1987) described beautiful experiments with a recursive description of networks which can be trained in the small and grown, and Waibel (1988) describes work in the modular combination of trained networks.

Connectionists are finally meeting a scientific goal for cognitive modeling which traditional AI itself has always lacked, namely the construction of theories embodied as computer programs which are both testable and repeatable. Meanwhile, the symbolic processing side of AI has wandered off into building systems of Brobdingnagian proportions (Lenat et al., 1986). Whether or not such large-scale software-engineering projects are even doable (Brooks, 1975; Cherniak, 1988; Parnas, 1985), AI's most ‘‘successful’’ programs, such as SHRDLU (Winograd, 1972) or BORIS (Dyer, 1983), are so complex that nobody can repeat these experiments to confirm or deny their findings or to build on previous results.

One of the hallmarks of real work in Artificial Intelligence is that theory-making must be backed up by program-building. Evidence of a working program is what distinguishes real AI from mere creative psychology. In one of the most brilliant turnabouts in scientific history, Minsky and Papert admit that neural networks are workable for toy problems, and therefore fit nicely into Minsky's grand unified theory of a Society of Mind (Minsky, 1986), a theory which purports that ‘‘mind is made up of a large number of components or ‘agents,’ each of which would operate on the scale of ... a toy problem’’ (p. 266). Unfortunately, while small and medium-sized neural networks are fully operational, nobody has ever seen a working Society of Mind program.

This is quite a humorous turn of events. The psychologists offer an implemented information-processing model, which the computer scientists reject in favor of a creative

---

[3] However, there are responsible practitioners who do care about tractability (Chandrasekaran, 1987).

psychological theory!

The remainder of the epilogue is, essentially, a criticism of certain aspects of modern connectionism, such as distributed representations (Hinton, 1984), from the perspective, and in the diffuse style, of Minsky's vast theory.

Conclusion

Perceptrons, and its authors, certainly have their places assured in history. The book absolutely needed to be reissued, as many copies were crusty tomes buried in the cybernetics sections of university libraries. And money definitely was at stake, in the high rolling game of academic publishing.

But this reviewer sees no scientific point to the new prologue and epilogue. They seem to be more of the same sort of stuff as the original introduction and conclusion -- harmful innuendo (i.e. "no better than exhaustive search") juxtaposed to, but not quite balanced by, whimsical praise:

> *On the positive side, we applaud those who bravely and romantically are empirically applying hill-climbing methods to many new domains for the first time. (p. 261).*

And nobody, not even the senior researchers of artificial intelligence, has a clue as to how the mind is constructed out of pieces, or how to software engineer, evolve, or back-propagate such a complex piece of machinery. All that we can hope for, today, is to build, from the middle out, a scientifically sound body of work on the mind. Minsky and Papert have created a lasting document with the mathematical core of *Perceptrons*. Whatever the authors' intention, the new prologue and epilog are not incontestable codicils.

References

BLOCK, H. D. (1970). A Review of *Perceptrons*. *Information and Control, 17*, 501-522.

BLUM, A. & RIVEST, R. L. (1988). Training a 3-Node Neural Network is NP-Complete. In *Proceedings of Institute of Electrical and Electronics Engineers conference on Neural Information Processing Systems*. Denver.

BROOKS, F. P. (1975). *The Mythical Man-Month*. Reading, MA: Addison-Wesley.

CHANDRASEKARAN, B. (1987). Towards a functional architecture for intelligence based on generic information processing tasks. In *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*. Milan, 1183-1192.

CHERNIAK, C. (1988). Undebuggability and cognitive science. *Communications of the Association for Computing Machinery, 31*, 402-412.

DREYFUS, H. L. & DREYFUS, S. E. (1988). Making a Mind versus Modeling the Brain: Artificial Intelligence Again at the Crossroads. *Daedalus, 117*.

DYER, M. G. (1983). *In Depth Understanding*. Cambridge: MIT Press.

HINTON, G. E. (1984). Distributed Representations. CMU-CS-84-157, Pittsburgh, PA: Carnegie-Mellon University, Computer Science Department.

HINTON, G. E. (1986). Learning Distributed Representations of Concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Amherst, MA, 1-12.

JUDD, S. (1987). Learning in Networks is hard. In *Proceedings of the First Institute of Electrical and Electronics Engineers Neural Network Conference*. San> Diego, II-685-692.

KOLEN, J. F. (1988). Faster learning through a probabilistic approximation algorithm. In *Proceedings of the Second Institute of Electrical and Electronics Engineers conference on Neural Networks*. San Diego, I-449-454.

LENAT, D., PRAKASH, M. & SHEPARD, M. (1986). CYC: Using common sense knowldeg to overcome brittleness and knowledge engineering bottlenecs. *Artificial Intelligence Magazine, VI*, 65-85.

MCCLELLAND, J. L., RUMELHART, D. E. & THE PDP RESEARCH GROUP (1986). *Parallel Distributed Processing: Experiments in the Microstructure of Cognition.* Cambridge: MIT Press.

MINSKY, M. (1986). *The Society of Mind*. New York: Simon & Schuster.

MJOLSNESS, E., SHARP, D. H. & ALPERT, B. K. (1987). Recursively Generated Neural Networks. In *Institute of Electrical and Electronics Engineers First International Conference on Neural Networks*. San Diego, III-165-171.

NEWELL, A. (1969). A step toward the understanding of information processes. *Science, 165*, 780-782.

PAPERT, S. (1988). One Artificial Intelligence or Many. *Daedalus, 117*, 1-14.

PARNAS, D. (1985). Software aspects of strategic defense systems. *American Scientist, 73*, 432-440.

PINKER, S. & PRINCE, A. (1988). On Language and Connectionism: Analysis of a parallel distributed processing model of language inquisition.. *Cognition, 28*, 73-193.

POLLACK, J. B. (1988). Recursive Auto-Associative Memory: Devising Compositional Distributed Representations. In *Proceedings of the Tenth Annual Conference of the Cognitive Science Society*. Montreal, 33-39.

POMERLEAU, D. A. (1988). ALVINN: An Autonomous Land Vehicle in a Neural Network. In *Proceedings of Institute of Electrical and Electronics Engineers conference on Neural Information Processing Systems*. Denver.

ROSENBLATT, F. (1962). *Principles of Neurodynamics*. New York: Spartan.

ROSENFELD, R. & TOURETZKY, D. (1988). Four capacity models for coarse-coded symbol memories. *Complex Systems, 2*, 463-484.

RUMELHART, D. E. & MCCLELLAND, J. L. (1986). On Learning the Past Tenses of English Verbs. In J. L. McClelland, D. E. Rumelhart & the PDP research Group, (Eds.), *Parallel Distributed Processing: Experiments in the Microstructure of Cognition*, Vol. 2. Cambridge: MIT Press.

RUMELHART, D. E., MCCLELLAND, J. L., & THE PDP RESEARCH GROUP (1986). *Parallel Distributed Processing: Experiments in the Microstructure of Cognition.* Cambridge: MIT Press.

RUMELHART, D. E., HINTON, G. & WILLIAMS, R. (1986). Learning Internal Representations through Error Propagation. In D. E. Rumelhart, J. L. McClelland & the PDP research Group, (Eds.), *Parallel Distributed Processing: Experiments in the Microstructure of Cognition*, Vol. 1. Cambridge: MIT Press.

RUMELHART, D. E. & ZIPSER, D. (1986). Feature Discovery by Competitive Learning. In D. E. Rumelhart, J. L. McClelland & the PDP research Group, (Eds.), *Parallel Distributed Processing: Experiments in the Microstructure of Cognition,* Vol. 1. Cambridge: MIT Press.

SAMUEL, A. L. (1967). Some studies in machine learning using the game of checkers. II - Recent Progress. *IBM Journal of Research & Development, 11*, 601-618.

SEARLE, J R. (1980). The intentionality of intention and action. *Cognitive Science, 4,* 47-70.

SEJNOWSKI, T. J. & ROSENBERG, C. R. (1987). Parallel Networks that Learn to Pronounce English Text. *Complex Systems, 1*, 145-168.

TOURETZKY, D. S. (1986). BoltzCONS: Reconciling connectionism with the recursive nature of stacks and trees. In *Proceedings of the 8th Annual Conference of the Cognitive Science Society*. Amherst, MA, 522-530.

TOURETZKY, D. S. & GEVA, S. (1987). A distributed connectionist representation for concept structures. In *Proceedings of the Ninth Annual Conference of the Cogntive Science Scoiety*. Seattle, 155-164.

WAIBEL, A. (1988). Modularity in neural networks for speech recognition. In *Proceedings of Institute of Electrical and Electronics Engineers conference on Neural Information Processing Systems*. Denver.

WALTZ, D. L. & POLLACK, J. B. (1985). Massively Parallel Parsing: A strongly interactive model of Natural Language Interpretation. *Cognitive Science, 9*, 51-74.

WINOGRAD, T. (1972). *Understanding natural language*. New York: Academic Press.