Neural Networks and Philosophy:

Why Aristotle was a connectionist.

Contemporary Problems:
Sue Becker

Module 3 Paper

Steve R. Howell
April 16th, 1999

I remember the moral philosophy course I took as an undergraduate at the University of Waterloo. It was only an elective for me, already confirmed as I was upon the path of psychology. Philosophy, however, is never wasted, always useful to an open mind. I found the course interesting, if apparently inapplicable. These years later, however, I find that things that I was taught in that class still resurface from time to time, usually in useful form to guide thinking about some problem in my current work.

Such was the case when I recently read Tim Rogers and Jay McClelland's 1999 paper on categorization with neural networks. In their opening arguments, they describe the issue of categorization with the classic syllogism "Socrates is a man, therefore Socrates is mortal". They illustrate that traditionally in cognitive science the method by which such reasoning takes place is considered to be categorization, categorization according to some process of rules. A rule exists in our minds with the information that 'all men are mortal'. Thus, when we are given the information that Socrates is a man, we can classify him as a man, and he inherits all the properties of 'man', including mortality. This use of explicit rules is what Rogers and McClelland seek to challenge with neural network models in the remainder of their paper, with some success as we shall see. However, what struck me at that point were thoughts of another Greek philosopher, the third generation philosophical descendant of Socrates, via Plato, one Macedonian named Aristotle.

In that undergraduate course in philosophy that I mentioned earlier, one of the major readings under discussion was the Nicomachean Ethics by Aristotle. In that work of moral philosophy, Aristotle examines the meaning of morality, virtue, and happiness. In moral philosophy generally, the investigation typically resolves around the search for what is 'good' or 'the good' in men's lives, the pursuit of which is the recipe for a good and meaningful life. Aristotle takes this in his writings to be synonymous with the concept of happiness, which is the ultimate end for which all intermediate actions are performed, even those whose results are to be found in the remote future or only in the aggregate. He concludes that moderation is the key to happiness, and thus that happiness is in fact what men mean by 'good'; problem solved!

The question that Aristotle spends much of the rest of that work addressing is how one is to know, in the course of one's life, which actions ('habits' as he puts it) will tend to lead to the 'good' of happiness. Since he does not believe it is possible to arrive at some explicit rule for the pursuit of the 'good' of happiness (the infinite variety of possible actions and generators of happiness precludes that), he arrives at what my psychology professors found to be a distinctly unsatisfying conclusion. Aristotle advises, in essence, that when in doubt about a life choice, one should find a ‘good’ man (or woman) and ask him or her what the path to good is, and make those paths one’s habits. Critics such as my professors found such logic circular and uninspiring. How, they asked, can this ‘good’ man know what is good? If he does know, how can he explain it to those he must advise? The good, happiness, is complex; it varies by situation and circumstance. The seekers after happiness would need to confer with this sage at every major point of choice in their lives, an impossible task. Thus did my professors dismiss Aristotle’s idea of ethics.

At the time, this dismissal bothered me intuitively, but I found no way to explain what it was that I disagreed with, nor how to attempt to demonstrate that Aristotle was on the right track after all. I now believe that the problem was that Aristotle was a 'connectionist' (perhaps the first), and my psychology professors were symbolists, logicians. They wanted rules that made sense a priori, without any messy ‘evolution’ of concepts over time. Aristotle by contrast with his ‘good man’ was thinking like a connectionist would today.

Examination of a few connectionist models will serve to illustrate this claim, including work on such various areas as categorization, reinforcement learning, and others. Consider the Rogers and McClelland example that I mentioned earlier. In that paper, the authors present a series of simulation experiments designed to investigate the process by which people acquire and use natural semantic knowledge. The authors contrast the traditional symbolic or logical explanation, involving categorization and categorization processes, with the connectionist or parallel distributed processing (PDP) approach. While both might loosely be considered ‘categorization’, the connectionist approach is quite different in processing, involving a more graded and distributed form of knowledge.

Rogers and McClelland are quick to point out that the operation of their model of semantic processing still leads to some of the same features as the categorization-based approach. However, they demonstrate that the classic difficulties of the categorization-based approach are reduced or eliminated in their model, and some paradoxes can be resolved. An important reason for this similar-yet different approach is related to the reasons why the categorization approach is traditionally forwarded. The proponents of such theories point to the power of categorization as a vehicle for generalization, as a process for computing similarity, and as an economical means of storage (Rogers & McClelland, 1999). However, these advantages are not limited to such categorization mechanisms. Importantly, they are also fundamental characteristics of the distributed representations learned by neural networks across a wide variety of domains. In the connectionist or neural net paradigm then, 'categorization', or its effects, arise not due to any pre-existing categories or processes of categorization. Rather, these observed effects are due to the gradual accumulation of semantic knowledge via the mappings between objects and their properties, as learned individually, in different occurrences, and aggregated via the mechanisms of the network.

This can perhaps be seen more clearly in Rogers and McClelland's analysis of the phenomenon of hierarchical structure. Their first simulation addressed the human data on progressive coarse-to-fine differentiation of concepts in development (Keil, 1979), and the reverse pattern of deterioration seen in various tasks in semantic dementia patients (e.g. Warrington, 1975). Their network was stopped at various points during training and the internal representations (states of the hidden units) examined. As might be expected by someone familiar with the processing of a PDP net, the hidden unit patterns start out relatively undifferentiated, but soon begin to activate differentially for different input patterns. Specifically, the broadest (or coarsest) distinctions are made first, then gradually finer and finer. Similar top-down distinctions are reached in other nets, the Elman model of language processing, for example. The simple recurrent networks used by Jeff Elman to investigate grammar learning learn first to divide nouns from verbs, then nouns into the different subclasses of nouns, verbs into transitive, intransitive, etc. (Elman, 1997). With the Rogers & McClelland net, the first distinction was animal vs. plant, then for example (within plant) flower vs. tree, etc.

The intuitively interesting thing about this straightforward task is the manner in which the network acquires this structure, regardless of the net we are considering. In the Elman network, the similarities were derived simply from co-occurrence statistics of words, while in the Rogers and McClelland case, apparent 'categorization' is due solely to the patterns of similarities between the semantic structure of the input patterns. The network acquires structure *automatically* through simple *exposure* to instances. The broadest distinctions are common to most if not all of those instances, thus the net has the most exposure to examples of this distinction (e.g. plant vs. animal). The lower-level (finer) distinctions will be common to only a smaller subset of the instances perceived by the network; thus it is exposed to fewer examples of that distinction, and should thus be slower to learn it. The finest distinctions, of course, occur in the fewest cases. That is, they occur only between very similar instances. Thus the finest distinctions will take the longest to learn.

Returning to the Elman networks, they incorporated an even more complex variable, that of the context of words. Internal representations of a given word were not the same in each occurrence of the word, but were different depending upon the temporal context, that is the words that came before it. In fact, the network learned these representations through the process of predicting the next word based on the context to that point, a process very plausible in childhood development according to both Elman and McClelland (1997;1994). This inclusion of context only added to the richness of meaning that could be inferred simply from the instances that the network was trained on.

The analogy of the above to the 'knowledge of the good' possessed by Aristotle's good man should be clear. The ability to categorize alternatives, into those that are 'good' and those that are not, could be based upon the classic sort of logical categorization, or upon the experientially acquired categorization performed by neural networks. With connectionist categorization, it is unnecessary to be able to state or conceive a 'rule' that would define the category. It was similarly unnecessary for Aristotle to abstract a rule defining the good. Rogers and McClelland's categorization net learned simply through exposure to many instances over time, thus providing a real, plausible mechanism to underlie the everyday concept of learning through experience. Categorization is learned solely on the basis of exposure to instances. Aristotle's good man was exposed to many choices in his life, some that led to good outcomes, some that led to bad. Those that led to good outcomes were reinforced by good fortune and remembered, defining a category by the pattern of their distributed characteristics. Then the man could, by consciously comparing any particular choice and its individual context to these 'categories', predict the goodness of it, that is the goodness of fit between this instance or example and the category 'good'.

It is important to note that several additional characteristics or assumptions of the connectionist approach are raised in this analogy. First is the issue of conscious processing. Mathis and Mozer (1996) discuss the issue of consciousness in relation to connectionism, and provide some valuable insight to this discussion. Second is the idea of reinforcement, and thus reinforcement learning. The entire field of reinforcement learning is relevant here, but a connectionist-oriented example can be found in Maclin and

Shavlik (1996). Third is the concept of prediction based on comparisons to distributed representations, prediction dependent upon context.

First, we can address the issue of conscious versus unconscious processing. It is important to note when discussing the connectionist approach that it does not purport to explain conscious processing, at least not in my experience with the field. Rather, connectionism deals with those vast areas of automatic processes that take place in our neural machinery, processes that simply exist in a black-box fashion. The results of this processing can affect our behavior directly, as in the case of more primitive or lower-level neural areas. As Mathis and Mozer point out, however, with higher cognitive processes the *results* of the process are the only things that enter consciousness. Examples of this might include such things as recalling a memory, or multiplying two numbers. We don’t know, and can't report, exactly how we do it. It just happens, our mental machinery takes care of it in automatic fashion, and we become consciously aware only of the results. Thus, the automatic, mechanistic approach of connectionism is in no way incompatible with consciousness. Certainly, in Rogers and McClelland's analysis of rule-based categorization versus connectionism categorization they did not deny the operation of rule-based classification. Indeed, if a rule is consciously perceptible, then people will often take advantage of it, and can thus consciously make a decision based on it. We can reason in the form of syllogisms; Socrates is a man, men are mortal, therefore Socrates is mortal. However, in less clearly defined cases, in all overlearned behaviors, and even in the processes underlying our retrieval of Socrate's name and the category *men*, we are making conscious use of the results of unconscious, connectionist machinery.

I suggested above that the good man had learned, post-hoc, which behaviors to consider leading to 'good' by the results that he observed for them. This can be considered reinforcement learning, a topic related to the connectionist paradigm even when not actually incorporating its methods. Back-propagation, the most common of the connectionist learning algorithms, is what is known as a supervised algorithm. That is, it receives from the environment, from a teaching signal, the exact difference between its output and what it *should* have output, and through incremental learning modifies its connections until that discrepancy is eliminated. Similarly, reinforcement learning is a type of supervised learning. Admittedly, the environment does not provide explicit calculation of how much one's behavior differs from the optimum for that situation. However, the simple consequences of one's actions are certainly perceived, either unconsciously in the case of lower level behaviors (e.g. don't touch a hot stove - pain!) or consciously in the case of higher-order ones ("well, that comment I just made was obviously taken as insulting, don’t say *that* again"). These consequences are a simpler form of supervision, where only the direction of the effect (positive or negative) and perhaps some idea as to magnitude (large insult versus small insult) are available. Furthermore, these signals, while originating from the environment as perceptions, are interpreted to a lesser or greater extent by the individual, who acts as his or her own supervisor (Freud's superego revisited!). Certainly in the case of higher order interpretations of the social environment, a great deal of conscious processing might have to go into determining the proper perception of a result, and its positive or negative value to the individual.

Maclin and Shavlik (1996) discuss an implementation of a reinforcement learning agent in an artificial environment. They use a neural network as the brain of their agent, which combines the two paradigms nicely. However, the innovation that Maclin and Shavlik add to the process of reinforcement learning is that of advice-taking. Rather than have the agent only able to learn from its direct experience in the environment, they include a mechanism to incorporate advice from an *observer* of the agent's performance. This is an interesting variant of the supervised learning approach, combining simple positive-negative reinforcement with more detailed, explicit discrepancy information. The observer can offer advice on any time as to which actions to take in which situations. The agent initially takes this information as reliable, but since it is incorporated into the same connectionist structure as its own learning, it can modify the advice or learn to ignore it with further experience.

Of course, Maclin and Shavlik are not the first to do this, but their technique of observer-initiated advice is relatively new, as compared to models in which the agent can request advice when it needs it. Both processes certainly operate in human development, or ethical development for that matter. In our 'good man' example, it is clear that most of the advice he would have to offer would have to be of the solicited kind, when approached by those seeking advice. However, there would be nothing to keep him from offering his sage advice to those who had not asked for it. To the extent that the advisee took such unsolicited advice, they would be operating like Maclin and Shavlik's reinforcement learners.

Thus the seeker after the path to goodness is not reliant upon their own connectionist neural machinery to learn categorization, but also considers (implicitly or explicitly) the positive or negative consequences of his or her actions, and may even make use of the advice of external experts. Of course, lacking advice of their own, how did those experts come to exist: pure chance? Perhaps.

Evolution is a concept usually confined to biology, but with the advent of genetic algorithms or genetic programming, we can examine virtual evolution. It is now known that the kind of end result, a trained agent or network, that can be created through either reinforcement learning or connectionist training, can also be achieved through evolution. Simply take a large population of agents, like the ones used by Maclin and Shavlik, and randomize the weights on their neural network's connections. Send each agent through the simulated environment. Most will do horribly, unable to accomplish the simplest tasks. A few however, will by chance manage to perform some function (typically measured as food finding or enemy avoidance) with some measure of success. Take those few, and 'breed' them; that is, combine their weights somehow, perhaps throwing in some random 'mutations'. Then send a new population of those agents through the environment. Repeat the procedure. Nolfi, Elman and Parisi (1994) have shown that such a procedure can produce agents every bit as capable as those trained via backpropagation or reinforcement learning, via a mechanism that essentially capitalizes on chance.

Thus, to return to our analogy, we have three different paths by which one may learn to be good. Direct training in ethics or morals, perhaps through story or tale (which in his "Republic" Plato knew enough to manipulate for the purposes of educating his future philosopher-kings) is a kind of learning by example, of distributed connectionist categorization though exposure to instances. Then, personal experience with good and bad, through the power of reinforcement learning, is self-interpreted and used to guide behavior. Finally, the reinforcement learning may incorporate the advice of an outside expert, Aristotle's good man, who can guide the learner more efficiently, having come by his knowledge either through the guidance of a prior 'good man' or by chance, in an evolutionary sort of way. Given that in philosophy, the 'good' is equated with the best, most efficient, and most productive life, the circle of interaction of these three influences should be ever-evolving towards perfection, as some (Whitehead, 1933) have argued is the course of civilization, past and future.

Thus my extended and somewhat facile example of the learning of morality through connectionism is neatly concluded. One might ask, however, the extent to which similar sorts of arguments might be extended to less restricted domains. Certainly I believe that almost anyone or any field can benefit from a consideration of the methods of connectionism and related work. Even without considering the approach in much depth, the ideas and concepts can provide an alternative to strictly logical and symbolic thinking in any domain. The connectionist viewpoint can serve other fields simply as a heuristic, a way of guiding thought and theory in new directions. Certainly we can apply it to philosophy, as just shown, and is not philosophy the original science, from which other ways of thought derive?

Furthermore, when considered in more depth, connectionist thought and methods can certainly provide new insight to the directly related, but broader, psychological issue of human development. Can we not, for example, consider similar mechanisms and arguments as were discussed above for the entire process of human learning and development, from childhood on? Random experiences, connectionist learning through exposure to instances, reinforcement learning, and advice-seeking and taking, all play a part in childhood development, and make us what we are. Nature versus nurture is an obsolete debate, as others have amply demonstrated (Elman et al., 1996); the two are closely intertwined, and set parameters for each other to operate within. More than just modeling parts of the process, connectionism can be the heuristic that guides our understanding of that process. Not only neural modelers should study connectionism, but also developmentalists, evolutionary psychologists, educators, and yes, philosophy professors.

References

- Elman, J.L., Bates, E.A., Johnson, M.H., Karmiloff-Smith, A., Parisi, D. and Plunkett, K. (1996). *Rethinking Innateness.* Boston: MIT Press
- Elman, J. (1995). Language as a Dynamical System, In Robert F. Port & T. van Gelder (Eds.) *Mind as Motion: Explorations in the Dynamics of Cognition.* Cambridge, MA: MIT Press, Pp. 195-223.
- Maclin, R. and Shavlik, J. (1996), "Creating advice-taking reinforcement learners", *Machine Learning* 22(1-3):251-281.
- Mathis, D. W. & Mozer, M.C. (1996). Conscious and unconscious perception: A computational theory. Proceedings of the 18th annual conference of the Cognitive Science Society, Cottrell, G. (Ed)
- McClelland, J.L. (1994), The Interaction of Nature and Nurture in Development: A Parallel Distributed Processing Perspective, in P. Bertelson, P. Eelen, G. D'Ydwewalle (Eds), *International Perspectives on Psychological Science*, *Volume 1: Leading Themes*, Erlbaum: Hillsdale, NJ.
- Melden, A. I., Ed. (1967). *Ethical theories: A book of readings.* Prentice Hall: Englewood, New Jersey.
- Nolfi, S., Elman, J.L., & Parisi, D. (1994). Learning and evolution in neural networks. *Adaptive Behavior*, 3:1, 5-28.
- Rogers, T. T., and McClelland, J.T. (In preparation). Semantics without Categorization.
- Warrington, E.K. (1975), "Selective impairment of semantic memory", Quarterly Journal of Experimental Psychology, 27:635-657.
- Whitehead, A. N. (1933). *Adventures of Ideas.* Cambridge: Cambridge University Press.