Neural Networks As a Path to Self-Awareness

Paul Werbos[1]

***Abstract*—There has been important new crossdisciplinary work using neural network mathematics to unify key issues in engineering, technology, psychology and neuroscience – and many opportunities to create a discrete revolution in science by pushing this work further. This strain of research has a natural link to clinical and subjective human experience – the "first person science" of the mind. This paper discusses why and how, and gives several examples of links between neural network models and key phenomena in human experience, such as Freud's "psychic energy," the role of traumatic experience, the interpretation of dreams and creativity and the cultivation of human potential and sanity in general, and the biological foundations of language.**

I. Introduction

In reviewing thousands of years of philosophy and religion, Nietzsche wrote [1]: "We are strangers to ourselves, we perceivers – we ourselves to ourselves; for this then is reason enough. We have never sought for ourselves – how, then, could it happen that some day we should find ourselves? … We must remain strangers to ourselves; we do not understand ourselves; we *must* mistake ourselves…" But perhaps this can change.

In referring to understanding ourselves, Nietzsche was not referring to whether we will ever have a big poster on our walls marking out every one of the neuronal transmitters or the genetic sequences which cause their expression. That can be useful stuff, but he was talking about something else. He was talking about the kind of understanding which helps us understand and decide what we are really trying to do or accomplish in life, and helps us understand the powers of our mind and the minds of others in a way which lets us be as effective as possible in whatever it is that really matters to us. Neural network research really has a crucial role to play in making this actually possible, more than has been possible anywhere in the past.

In this paper, I will give a few concrete examples of how this works, and of things I have learned. But first I will address a few predictable questions: (1) Why should we care?; (2) Isn't this a ridiculous exaggeration of what the neural network field is actually doing?; Doesn't psychology and cognitive science already take care of this? (3) Are you another one of those guys like Dennett and Nietzsche and the transhumanists bent on discrediting the human soul itself, along with what has been learned in thousands of years of human culture?

[1] Paul Werbos, werbos@ieee.org, is with NSF, with CLION, and with IntControl, but this paper represents personal views only, not the views of any of those organizations.

A. *Why Should We Care?*

Why should we care about knowing what our true ultimate values are, or about how to accomplish them more effectively?

At some level, it is weird that people even ask this question. It is an example of the common pathology called "denial." [2] Every day we see plenty of expensive research projects and studies and even government agencies which end up being a total waste of time and money, all because people did not ask the right question or formulate their goals in a rational enough way. In addressing complex business or strategy decisions [3-6], it is crucial to make the effort to define one's values as clearly and precisely as possible, as a kind of utility function. Without some degree of focus and conscious effort towards goals, we know from experience that it is hard to accomplish anything.

Or is it?

For example, could we make it through life just as well by following simple, rigid stimulus-response rules? Animal psychologists have sometimes proposed that all behavior can be described in terms of inborn fixed stimulus-response rules. But in fact, we as mammals (and many other organisms [7]) are not so limited. Evolution discovered long ago that we can get better results if we keep changing our behavior based on learning, learning which changes stimulus-response patterns in order to get better results. This in turn requires that we do have some kind of unified system, at the end of the day, to evaluate which results are better than which other results. It basically requires some kind of inborn "reward" or "pleasure/pain" or utility or primary motivation system[8]. By this logic – *we already do care* about the results of our actions and decisions, whether we admit it or not.

But should we admit it? When we start to articulate what we really care about, it gets to be like looking at ourselves in a mirror. It is understandable that many people fear what they might see in the mirror if they look too hard. It is even more understandable that they do not trust their initial understanding of what they think they see in the mirror, and do not want to put too much weight on it. A key role of neural network research is to help make it more possible to understand and to keep looking in the mirror.

When people choose not to articulate their true goals, and reason about them in words and mathematics, they basically choose to live the core of their lives at a subsymbolic level, like a mouse or a rat. If you had a conversation with a mouse (or a certain kind of bean-counter), it might well ask: "Who needs a human type brain anyway? What good is it? Just how much cheese do you get per gram of white matter? Can I eat it?" But again, experience has shown that symbolic intelligence does have its uses. It is simply a matter of

following nature for humans to try to learn to get full use of that faculty, along with other mental faculties we are born with but must learn to use.

Finally, there have been some philosophers who would assert: "We agree with the idea of ultimate goals or ethics, but we don't care about human subjective experience or feelings. We believe that goals should be deduced from pure and perfect logic, or objective science." Yet modern logic makes it perfectly clear that this is impossible. Starting from axioms which do not contain value words like "good" or "should," it is impossible for valid logic to result in any kind of theorem which contains such words. Even starting from experimental data which is not labeled with values, it is impossible to learn values. Objective science simply cannot answer the question "what SHOULD I do?" But it *can* help us answer the question: "What *would* **I** do 'if I were wise'? What pattern of goals would really satisfy me, as a whole human, considering all aspects?" The latter is answerable, because it contains the word "I," which we can start to understand by using science to study ourselves in the mirror. In the mirror, we can easily see that symbolic reasoning is at best like the trunk of a tree, which will dry out and become totally dead if it is not firmly grounded in its subsymbolic roots; full use of such a brain involves the integration or harmony of the (learned) symbolic reasoning with the powerful subsymbolic intelligence, which is always in charge and always plays a key role.

Politically minded people (like many philosophers and even more religious rulers) sometimes prefer that people not spend too much energy understanding what their personal goals are. Sometimes they even prefer that their subjects not learn to read, lest they become more powerful and threaten the throne. They prefer that all attention be focused on collective political goals, imposed from above. But societies can often be more effective if they encourage their citizens to "be all they can be" as individuals, and use mechanisms like conscious social contracts and Pareto optimal interaction to work together. Humanity today is facing several serious threats to its very survival; a quantum jump in human capability, self-awareness and potential may be risky, but also essential to our ability to rise to those challenges. (Or in other words, there are a lot of crazy folks out there who seem to be on a path to getting us all killed, if we don't find a way to inject more sanity, fast.)

Many of the points discussed in this section seem rather obvious, or even trivial, from an objective point of view, from what we see when we look at ourselves in the mirror. Yet for centuries and centuries, people and societies have gone through great struggles, as they did not really see or assimilate what should have been obvious. As an example, the previous paragraph is basically a quick logical summary of the great cultural struggle between the followers of Meng Tzu (aka Mencius) and the followers of Mo Tzu and of Qin, leading up to the creation of the Han dynasty, the first truly effective government of all of Han China. [35]. Everything in this paper takes a position generally consistent with the fundamental viewpoints of Meng Tzu. Likewise, the concept of utility function here is basically just a clearer, more modern version of Aristotle's concept of telos and of the philosophy of John Stuart Mill in its more adult form [36].

B. Neural Networks And Psychology

But what do neural networks have to do with this kind of self-understanding?

If Congress or DARPA were to drop $300 million into a new all-encompassing interagency initiative on neural networks, or on neuroscience, it would probably do more harm than good to the goal of self-understanding. Historically, psychology or cognitive science have been the disciplines which focus on scientific understanding of the mind. Psychology and cognitive science have certainly made important contributions relevant to this goal [9-13], but $300 million of new money to psychology in general would probably not have transformative benefits either. Neural networks, psychology, neuroscience, and cognitive science are all very large fields, which deserve strong funding, but do not really focus on the specific new opportunities which are most important to a new level of self-awareness.

In essence, these cross-cutting new opportunities involve a focus on three key questions:

(1) How can we model and replicate the high-level mathematical principles which allow the brain and mind to learn over time to do better and better in maximizing whatever utility function comes to it, in whatever form, from the primary motivation system? (The mathematics of natural intelligence.)

(2) What kinds of primary motivation system have actually evolved in nature, and what are the large-scale implications?

(3) How can we develop some kind of two-way synergy between (1) and (2) and our direct experience of life, and the realization of human potential?

The immediate grand challenge to hard science is with the first question. In [14], I argued that we now have all the elements we need for a scientific revolution in understanding intelligence in the brain, as fundamental and mathematical as the kind of revolution which Newton carried out in the 1600's. (Though could it be that today's Western society is more conservative and resistant to paradigm shifts than it was then?) In 1997, in developing the NSF initiative on Learning and Intelligent Systems [15], some of us posed the challenge as follows: (a) We know that vertebrate brains are basically networks of simple processing units like neurons, and we can define a "mathematical neural network" as any dynamical system driven by learning with that general kind of parallelism and such; (b) the challenge is to use a COMBINATION of experiments in engineering and technology, together with experiments in neuroscience and behavior, to filter through the vast space of possible mathematical neural networks and locate at least one universal learning model which satisfies both filters.

I also remember telling Karl Pribram that someone high-up then interceded: "Why only neural networks? Why not something else?" Karl's eyes lit up, and he was probably hoping they would make room for concepts like field effects and even laser, quantum kinds of effects in the brain, for

which he felt there was substantial empirical evidence [37]. But then I said more: "The higher-up then explained he wanted the geology division to be able to participate." "What was in *their* brains?" Karl may have wondered, with great dismay. But in fact, Pribram's concepts in [37] can be fully accommodated within the larger field of possible types of mathematical neural networks [16]. See the appendix for a brief history of some of the better known forms of mathematical neural network.

More recently, the Engineering Directorate of NSF supported a more focused one-time crossdisciplinary effort to follow up on this opportunity [17]. As with the first year of LIS, it produced tantalizing initial results (like the new world records on object, phoneme and text recognitions by LeCun [18,19]) – but larger and more sustained funding would be needed to do full justice to the area. I have done my best through the years, on limited personal time and with help from CLION, to fill in many of the details of this opportunity [20-22], spelling out the mathematical principles which make it attainable.

But what about question 2 (motivation) and the issue of direct, first-person subjective experience? Some of the basics of human motivation are relatively obvious [8], and will become easier to study when we better understand the intelligence side. Most of what we call "emotions" are actually part of the intelligence side – flows of affect or valuation or secondary reinforcement which emerge as part of learning. I have done my best to make the connections to the world of subjective experience in my own work, but have only injected a few important examples into the published literature (and a few more into www.werbos.com).

C. *What About the Soul and Traditional Human Cultures?*

In a session on subjective experience, it may be appropriate to recall some. After all, direct experience is the empirical foundation of first-person science.

Early in 1967, as a senior at Harvard, I had fully assimilated all the logic above. My views about the brain were just as clear, mundane and hard core as those of folks like Dennett, and my views of religion and philosophy were similar to those of Nietzsche. Because half my family were a kind of Druid Irish Catholic, I had had some personal experience which put some stress on my convictions [23], but I resolutely stuck to the logic of the situation – especially to Hebb's Bayesian analysis of parapsychology in [24].

But then consider the context. I had also fully assimilated the understanding that maximum effectiveness and rationality come from a full integration of symbolic and subsymbolic intelligence. In the tradition of German existentialism (the other half of my family), I had decided that I would do whatever I felt like ever more – but that I needed to be somewhat more open to my feelings. Usui has told me that Zen Masters complain about klutzy all symbolic left brain thinkers who do not even bother to notice whether the sushi they are eating is the very best or rotten; in a similar way, I decided to really notice the feelings generated by different foods in the Adams House cafeteria, and to really welcome and feel the music I would play in the background while studying, thinking or in bed.

Since I was deeply, whole-heartedly focused on trying to understanding intelligence in the brain, I naturally was drawn to any source of information at all that could help unravel the mystery. I spent a lot of time trying to make sense of the wiring diagrams and behavior one could glean from real systems-level work in neuroscience such as [25] and [26]. But a few times, I couldn't help trying to "cheat" by looking down with great intensity on the flow of ideas and such in my own brain, as best I could, using some knowledge of anatomy to guide me, with that music in the background.

Those who have tracked traditional schools of human development from any part of the world would not be so surprised as I was at some of the unintended, unplanned outcomes. Like it or not, I had no choice but to change my worldview, and ask many questions I had not gotten so deep into before [27]. I apologize to those who disagree – but I would also plead that they be tolerant of those of us who have defected to agreeing on a few points with the majority of the other people on earth. There is a place for some diversity on these issues – and, indeed, if Western civilization should try to fight too hard against such diversity, it might endanger its own existence. (Living in Virginia, I have also visited George Washington's meditation room, and learned a lot about many of the past strengths of the US – not good things to shut down.)

At that point, I worked very hard to reconcile the obvious cognitive dissonance I had fallen into – and, with my Germanic genes, I really do not feel comfortable with cognitive dissonance. (Nor does my Russian wife; low tolerance of cognitive dissonance is known to have a major genetic component, but the distribution of those genes is complicated.) I was relieved to learn that the vast majority of successful PhDs in the US have had to confront similar types of challenge in their own experience [28], and have been quiet mainly because of social taboos and not knowing how to begin to put things into rational order after that. I also re-examined what we really know about the underlying laws of physics and what they permit [29]. In the end, I conclude that mathematics, physics and experience do allow for a level of intelligence as far above the isolated human brain as the mouse is above the reptile or more, based in part on quantum principles but, even more, on some kind of collective intelligence effects, not unlike what has been described by Carl Jung and many others in human history. Just as understanding the mouse is a key prerequisite to understanding the human brain and what it can do with symbols, it is also a key prerequisite, in my view and my experience, to making some kind of sense of what would otherwise be a bewildering realm of experience enough to drive one crazy if one did not have such a mathematical foundation.

But at first in 1967 – my immediate thought was "How could I possibly hope to model or understand this kind of stuff if I do not have enough data?" So I read rather widely across all potentially relevant literature, not paying so much attention to the theories, but trying to extract whatever

underlying experience or empirical data could be found, and trying to find a way to get more primary direct information as well. This did entail probing as deeply as possible to learn from cultures and sources from all over the earth.

And so … many, many years later I found myself in the opulent house of a neuroscientist scion of the family which leads one of the oldest and most powerful orders of Sufis from central Asia. We had many interesting discussions, but we did come back to this issue of neural network models. "Why do we need mathematics for the highest level of consciousness?" he asked. As a kind of Pythagorean, I argued that mathematics is really just the language of precision, a language which allows us to express and visualize things far more complex than can be captured in words. It is basically a matter of trying to understand our whole realm of experience as completely as we can, as much as we can. It is far too early to give up on that approach, as no one on earth has ever encountered anything which is not consistent with that general approach – though of course we still need to be open to direct experience in a practical way at the same time.

II. A Few Examples of Neural/Experience Links

A. Freud, Qi and Backpropagation

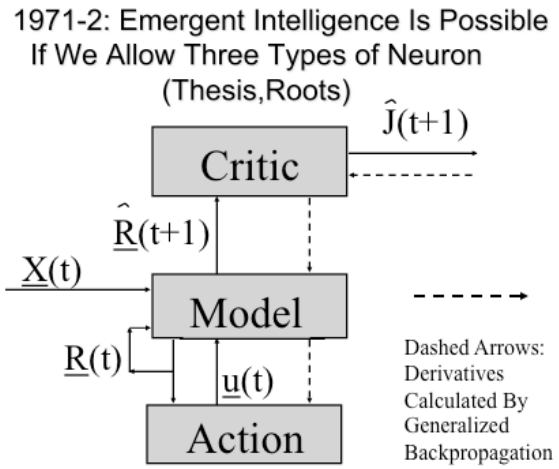

Fig. 1. Where Backpropagation Came From

The general form of backpropagation was first published in my 1974 Harvard PhD thesis (reprinted in [14]), the ancestor of the later "second generation adjoint" methods [30] for calculating derivatives, and far more general than earlier methods from Jacobsen and Mayne which focused on impulse matrices from time t to time t+1 without considering the case of neural networks or sparse networks in general. (Ironically, some computer scientists have recently popularized the rumor that it was developed by a person who cited the work of Jacobsen and Mayne, but who did not believe in the 1970's that backpropagation could even work at all, or yield accurate derivatives, for neural networks.)

Figure 1 gives the initial version of how backpropagation fits into a general intelligent system, as in the thesis prospectus for Harvard. The idea was to use the new algorithm to do local calculations to calculate the derivatives of "J*" with respect to all the variables in a "brain" made up of three interacting neural networks. (See [20] for more details.) In subjective terms, J represents a kind of learned measure of well-being, like what Skinner would call total secondary reinforcement or learned reward. The derivatives of J with respect to any variable R_i in our field of perception represent the *value* of increasing that value, to the organism. The model in Figure 1 is actually just a mathematical version of Freud's model of psychodynamics, where the derivative of R_i represents what Freud called the "cathexis" (or affect) or emotional charge or emotional energy attached to the object which R_i represents. In other words, I came up with backpropagation by *not* just laughing at Freud's "nonscientific" model, but by translating it into mathematics and showing that it works.

More concretely, Freud said that "if A causes B, a forward association or axon develops from A to B; and then, if there is an emotional charge on B, that energy flows backwards, to put a charge in A, proportional to the charge on B and to the strength of the forwards association from A to B." That's exactly what backpropagation does. Chronologically, I translated Freud into a way to calculate derivatives across a network of simple neurons (which Harvard simply did not believe), and then proved the more general chain rule for ordered derivatives to prove it and make it more powerful (and to graduate).

Freud's term "psychic energy" for this flow really captures the subjective feeling of this subjective reality, which Freud documents many, many times over in his works. (Though of course, it is not conserved like the energy operators of physics. It is a computational flow, however implemented.) But in my view, any really strong collective intelligence would have to be held together by the same kind of thing, propagated over a more complicated network topology, but still the same basic thing. And indeed, almost every major deep culture on earth has a term for the same kind of "psychic energy" at another level – like "qi" or "prana" or "charisma." What's more, several different types of derivatives (like derivatives of J versus derivatives of error) need to be propagated, giving rise to different kinds of mental energy. Sensitivity to these flows, to the fact that they are not conserved, and to the mathematics of the factors which govern their flow, is of great importance, in my view and my experience.

The challenge of how to respond as well as possible to different levels of feedback flows (and to deploy feedback) is a perpetual challenge at many, many levels of human experience. It calls for sensitivity, for active efforts to understand the feelings, motives and strategic considerations behind the flows, for systems of active dialogue and exchange, and for many applications of the concept of Pareto optimality, such as efficient market design and the "alchemical marriage." This deserves a lot more explanation and elaboration, but goes beyond the scope of this paper.

B. Mirror Neurons and Vicarious Experience

Back in the days of B.F. Skinner, there were many debates about whether human brains are basically equivalent to rat

brains, in the type of learning. Some psychologists argued that humans are different, because humans – unlike rats – can learn from the experience of other humans, which they assimilate as if their own. Because the approach then was not so mathematical and engineering-oriented as modern neural networks, their arguments were too fuzzy to prevail in an era when psychology had deep commitments to the Skinner orthodoxy. (Not that we are free from nasty orthodoxies today!)

Later, as the ideology of humans as pure symbolic intelligence started to spread, I proposed a different view of human language [chapter 10 of [14], [20]]. Because the higher part of the human brain is 99% parallel to structures in the mouse brain, and because humans are not born with reliable instincts to perform logical reasoning (let alone correct logical reasoning), I proposed that our ability to reason with words and with numbers is on an equal cognitive footing, as something learned, and possible because of a much subtler change in brain structure. In effect, I predicted "mirror" neurons as the first stage of this process even before they were actually found by Rizzolatti (and later interpreted by Arbib). When we think in terms of memory-based neural network learning, it becomes very clear and graphic that training the subsymbolic neural networks to a larger database of past experience leads to dramatically different results from a database of one's own experience only. Mirror neurons prove that we (and chimps) do possess this important capability, which rats and mice do not. This by itself is already enough to establish that there is a deep fundamental basis for *empathy,* one of the most fundamental and important faculties of human minds. It also helps us to understand our use of language better – but that topic is beyond the scope of this paper.

One interesting further implication: consider how dreams are basically simulations [20,31] grounded in things we have experienced or goals we have thought about. If our models say that the database this is operating form is actually a database including both our own experience and the experience of others, we would naturally predict that many of our dreams are actually based on *someone else* as protagonist! I have often wished I knew a clinical psychologist or psychiatrist willing to co-author a paper on how this can be seen in the interpretation of dreams in clinical practice. But at least I can say I have often found it amazingly instructive when I have taken the effort to remember some of my own dreams in the early morning – especially those dreams which would not make as much sense with me as protagonist.

Sometime the neural network field and the field of human society seem like one vast forest of mirrors, in which symmetry principles of all kinds are the underlying key to making complex things work. (The neoPlatonist expression "as above, so below" refers to a subset of these symmetries, which they relate in turn to a diverse set of life experience.)

C. Traumatic and Euphoric Memories

Neural network research also sheds new light on another key idea of Freud, rooted in subjective and clinical experience – the idea of traumatic memories.

Freud argued that many nervous breakdowns and phobias are due to the influence of undigested traumatic memories, dating back to childhood. But with proper methods, he argued that people could relive and digest such memories "into the ego," and bring the patient back to sanity. This is another key insight which many have considered unscientific or incomprehensible – but can be understood more clearly than in Freud's time, with the help of neural network approaches.

The key issue here in neural networks is the balance between fats real-time memory formation and slow generalization, which requires a balance between learning in real-time and learning from memory. Strangely enough, few engineers have followed up on this issue, even though the reconciliation of good global generalization (which requires good global modeling via good universal function approximators) and fast real-time learning (which is essentially possible only for memorization networks, when there is a large number of variables) is a key challenge to the engineering field. Many years ago [32] I proposed a solution to this problem, which I called "syncretism." Perhaps the more modern terminology and description in [22] will help remedy this gap.

The key design here again follows Freud – the use of a memorization type network (corresponding to Feud's "id") and a global model (which fits use of the word "ego" in *this particular context*), used together to make predictions. The brain cannot afford to revisit every memory of each past event, in real time; thus it is forced to use a *combination* of the ego with a more associative kind of predictor specifically to approximate effects which are not yet understood. If the "ego" has become more powerful in the time since the memory was last relived or revisited (for purposes of memory-based adaptation of the ego), then after a new reliving the memory will lose a lot of its ability to change one's expectations form what the ego presents. Curiously enough, the scientologists have built an entire religion (with many less scientific elements) on this very valid therapeutic principles, without proper acknowledgement of Freud.

The "syncretism" model makes a number of other interesting predictions, which do seem to fit my experience: (1) the aberrating effect can be euphoric (irrational optimism) or even neutral just as much, and with just as much risk, as traumatic; (2) reliving can actually make the neurosis worse, if the ego has not learned in the meantime how to really make sense of the earlier memory; (3) a healthy brain would not mature towards al-ego, but would keep accumulating new unexplained memory in new realms, as the circle of the known grows but the circle of the unknown also grows.

D. Bootstrapping to Enhance Awareness

Neuroscience has long studied issues like what happens to kittens reared in darkness, who lose the "access" to visual inputs and never learn to see, even in the light, unless they receive special stimulation like bicuculline. In some cases, this may occur because the inputs never get to the thalamus or to the olfactory bulb, the two gates to normal awareness. In other cases, it may occur because of "salience" problems.

In either case, our models suggest that "learning to predict" [20,22] is the main accessible driver of this system. Thus real-time experience in which one needs to pay attention to new inputs in order to predict and keep tracking something one is already paying attention to can be one useful tool in overcoming such limitations. Of course, manipulation of salience signals in one's own brain and mind can also help.

E. Escaping From Local Minima

Many theorists, disconnected from subjective reality, have argued that real biological neural networks could not possibly be using backpropagation or optimization as part of their learning, because systems like that can get caught in local optima. Systems like that could never be a good model of human intelligence, they say, because they cannot solve NP-hard problems like playing a perfect game of chess.

This is somewhat odd, because most of us know that humans do not play a perfect game of chess. If any design for a brain within the available hardware constraints (e.g. no more than a few trillion processors) could learn to play a perfect game of chess and go and all the rest, why would nature not have evolved such a capability? The obvious explanation is that it can't be done, even with trillions of processors. Approximation is an inevitable fact of real life. Systems which live in truly complex nonlinear environments simply cannot find the exact, global perfect optimum in a realistic amount of time. By thinking about this, we can immediately see: (1) we too must be caught in local minima all the time; and (2) nature must have evolved multiple levels of mechanism to reduce the resulting damage.

When I mentioned this at an IJCNN conference a few years ago, I was amazed that one person got up and said: "I am not in any kind of local minimum. I am absolutely perfect." I do hope that thinking it over would help him to realize that this was not write the whole truth, and that every one of us has many opportunities to do far better in life.

In fact, if you think about our cousins the chimpanzees – those folks are very intelligent, and even have mirror neurons, but if a human lived like that, we would agree that he or she really is caught in a rut. Humans got out of that same rut, even though we do not have a full mastery of language and symbolic reasoning, because we have just enough ability to learn and use language and mathematics to build technologies a bit more creative than the chimpanzees normally display. The chimpanzees and mice are in turn far more creative in finding new "basins of attraction" than reptiles, I would claim, because of certain functions of a stochastic layer of the cerebral cortex [20] which learns how to come up with new options in a complex way, far more sophisticated than simple reliance on the fixed kinds of methods used in traditional evolutionary computing.

Awareness of these creativity functions – in our personal lives , our social lives and in our social organizations – can have many potential benefits (e.g., see [5]).

F. From Mirror Neurons to Dance to Language

Mirror neurons by themselves are not enough to explain how humans learn language and symbolic reasoning far better than other apes do. Why are they not enough?

The answer [14] seems rather obvious from a neural network perspective. The mirror neurons as observed by Rizzolatti would make it possible for a monkey or a chimpanzee to learn from an augmented memory, augmenting the monkey's own experience with the experience of other apes *which it observed directly*. That was a big step in evolution.

The next step is something I recall very vividly from anthropology films shown by DeVore of Harvard, in classes on primate behavior, which basically fill in details of the larger framework developed by E.O. Wilson [7]. One of these films depicted the lives of well-preserved ancient African hunter-gatherer cultures (Kalahari bushmen), whose survival depended on big hunts conducted by males over many days in the bush.

After those hunts, the hunting party would *dance out* the story of what they just experienced. Because hunting is a complex and difficult activity, in which errors can be costly, it was important for other members of the tribe, like the youth, to learn the lessons of what worked and what did not. It was important for them to learn from experience which they were not present to observe directly. Thus, *building upon* the mirror neuron structure, early humans evolved the ability to augment their personal, subjective memory database still further, by including *reconstructed* experience, and by engaging in acts of communication (first of all like the Busman dance) which make this a two-way exchange, from memory to memory in effect. In that film, many members of the tribe went into a kind of hypnotic trance as they vicariously lived what others were dancing out; this fits well with the observation that humans are the only species on earth capable of true hypnosis[33].

This leads to the following hypothesis: until a few thousand years ago, human language was basically just a "dance of sounds." It was just a refined version of the bushman dance, used to dance out a kind of "word movie" of human experience. Just as human memory may contain prototypes or archetypes which summarize multiple observations in a single abstract example, these word movies evolved to convey that kind of abstract experience as well. The hard-wired mechanisms for language which are born into our brains are mechanisms for sharing concrete or abstract experience. This fits well with the observations by Sapir [34] that early languages followed a "grammar" more like a word movie than like a formal proposition in language.

Max Weber and others long ago noted that modern languages have been shaped very heavily by changes in culture in historic times. For example, the specific concept of "logic" developed by Socrates and Plato proliferated the concept of a "proposition" as a kind of object in some kind of space of abstraction. The idea that language should consist of "complete sentences" or well-defined logical propositions emanated from there, and has been enforced with great vigor by many school teachers doing their best to overcome the less restrictive inborn predispositions of the human creatures in their care. The use of the word "that" as a subordinating conjunction in English is an especially beautiful example of how the language can bracket a

proposition (clause), so that the proposition itself may be discussed as a kind of object. These mechanisms (not much older and no more biological than the concept of theorem ala Euclid) then empowered us to apply our natural skills in strategic thinking and planning to the realm of symbolic reasoning itself, both verbal and mathematical.

G. From Language to Self-Awareness

Once humans began to rely more heavily on propositional logic, in recent millennia, new dynamics started to come into play, not so directly related to the brain as such.

In the early stages of language use, human utterances were basically just "articulations" – a kind of *translation* of experience or generalized experience from the subsymbolic level to words or mathematics. But more and more, as we engaged in logical reasoning within the realm of symbols, we relied more and more on some kind of *axioms,* on certain privileged propositions and ways of reasoning, which became ever more powerful in our lives. Because all of this is relatively new and complex, natural selection has not yet had time to keep it from going amok; we therefore depend on our own subsymbolic learning abilities (both personal and social) to keep it from going fundamentally awry.

When the fundamental axioms that we use in symbolic reasoning directly express the intelligence and feelings in our subsymbolic mind, the two can support each other and achieve maximum harmony and effectiveness. But when the organism relies on a set of axioms which conflict with the subsymbolic mind, there will always be circumstances which set off a very fundamental conflict and instability, and potential breakdown. When such a conflict looms, there is often a kind of race in time – which will die first, the illusions or the person who harbors them? Perhaps we have reached a point in history where this also applies to humanity as a whole.

REFERENCES

[1] Friedrich Nietzsche, *The Genealogy of Morals.* William A. Haussmann translator, 1897 (Google Books)

[2] George E. Valliant, *Aging Well: Surprising Guideposts to a Happier Life from the Landmark Harvard Study of Adult Development*. Little, Brown and Company, 2003

[3] H.Raiffa*, Decision Analysis*, Addison-Wesley, 1968.

[4] Von Neumann J.Von Neumann and O.Morgenstern, *The Theory of Games and Economic Behavior*,Princeton NJ: Princeton U. Press, 1953.

[5] Hitch, Charles Johnson, 1967, *Economics of defense in the nuclear age*. Ann Arbor, Michigan: University of Michigan, 1967, www.lib.umich.edu

[6] P. Werbos, Strategic Thinking for Leadership in Science &Technology, in W. Bainbridge (ed), Leadership in Science & Technology, Sage, in press

[7] E.O. Wilson, Sociobiology: The New Synthesis, Twenty-Fifth Anniversary Edition, Harvard U. Press, 2000

[8] P.Werbos, Values, Goals and Utility in an Engineering-Based Theory of Mammalian Intelligence, in Karl H.Pribram, ed., *Brain and Values*, Erlbaum: Hillsdale, NJ, 1998.

[9] William James, The Principles of Psychology, Cosimo Classics, 2007

[10] Miller GA, Galanter EH and Pribram K, *Plans and the Structure of Behavior.* Holt, Rinehart and Winston, 1960

[11] D.Rumelhart and J.McClelland, *Parallel Distributed Processing*, MIT Press, 1986..

[12] D. Yankelovich & William Barrett *Ego and instinct: The psychoanalytic view of human nature--revised*, Random House 1970

[13] Simon, H.A. *The Sciences of the Artificial (3rd ed.)*. MIT Press, 1996

[14] P.Werbos, *The Roots of Backpropagation: From Ordered Derivatives to Neural Networks and Political Forecasting*, Wiley, 1994

[15] http://www.nsf.gov/pubs/stis1996/nsf9718/nsf9718.txt?org=NSF

[16] P.Werbos, Quantum theory and neural systems: Alternative approaches and a new design. In K. Pribram ed., *Rethinking Neural Networks: Quantum Fields and Biological Evidence*, Erlbaum, 1993.

[17] http://www.nsf.gov/pubs/2007/nsf07579/nsf07579.htm

[18] LeCun 1 Koray Kavukcuoglu, Pierre Sermanet, Y-Lan Boureau, Karol Gregor, Michaël Mathieu and Yann LeCun: Learning Convolutional Feature Hierachies for Visual Recognition, *Advances in Neural Information Processing Systems (NIPS 2010), 2010*

[19] Yann LeCun, Koray Kavukvuoglu and Clément Farabet: Convolutional Networks and Applications in Vision, *Proc. International Symposium on Circuits and Systems (ISCAS'10), IEEE, 2010*

[20] P.Werbos, Intelligence in the Brain: A theory of how it works and how to build it, *Neural Networks*, Vol. 22, Issue 3, April 2009, Pages 200-212

[21] IEEE Trans Kozma R. Ilin, R. Kozma, P.J. Werbos, "Beyond backpropagation & feedforward models: A practical training tool for more efficient universal approximator," IEEE Trans. Neur. Netw. 19(3), pp. 929-937, 2008.

[22] P. Werbos, Mathematical Foundations of Prediction Under Complexity, Erdos Lectures/Conference 2010, http://www.werbos.com/Neural/Erdos_talk_Werbos_final.pdf

[23] http://drpauljohn.blogspot.com/2010/04/true-ghost-story.html

[24] D.O.Hebb, *The Organization of Behavior*, Wiley, 1949.

[25] Francis O. Schmitt, The Neurosciences: (First and Second) Study Program, Rockefeller university Press, ----- and 1970

[26] Sebastian Peter Grossman, Textbook of Physiological Psychology, Wiley, 1967

[27] P. Werbos, Space, Ideology and the Soul: A Personal Journey, in Bob Krone ed., Beyond Earth, Apogee Books, 2006. http://www.werbos.com/Space_personal_Werbos.htm

[28] Greeley, A. M., & McCready, W. C. (1975). Are we a nation of mystics? *New York Times Magazine*, Jan. 26, 1975

[29] Werbos, P. 2008. Bell's Theorem, Many Worlds and Backwards-Time Physics: Not Just a Matter of Interpretation, *Intl J Theoretical Physics*, e-pub date 2 April 2008. http://arxiv.org/abs/0801.1234

[30] P. Werbos, Backwards differentiation in AD and Neural Nets: Past Links and New Opportunities. In H. Martin Bucker, George Corliss, Paul Hovland, Uwe Naumann & Boyana Norris (eds), *Automatic Differentiation: Applications, Theory and Implementations*, Springer, New York, 2005

[31] Stephen LaBerge, Lucid Dreaming, Sounds True, 2006

[32] P.Werbos, Advanced forecasting for global crisis warning and models of intelligence, *General Systems Yearbook*, 1977 issue

[33] George A. Estabrooks, *Hypnotism*, Plume, 1959.

[34] Edward Sapir, Language: An Introduction to the Study of Speech, Harcourt-Brace, 1921 (Google Books)

[35] Fung, Yu-Lan, *The spirit of Chinese philosophy*, Beacon Press, 1967

[36] Mark S. Micale, *Hysterical Men: The Hidden History of Male Nervous Illness*, Harvard U. Press, 2008

[37] K.Pribram, *Brain and Perception: Holonomy and Structure in Figural Processing*, Erlbaum 1991.

[38] M. Minsky & S. Papert, *Perceptrons: An Introduction to Computational Geometry*, MIT Press, 1969

[39] Anderson, J., and E.Rosenfeld, eds, *Talking Nets*, MIT Press, 1998

[40] P.Werbos, Applications of advances in nonlinear sensitivity analysis, in R.Drenick & F. Kozin (eds), *System Modeling and Optimization: Proc. IFIP Conf. (1981)*,Springer 1992; reprinted in [14].

[41] Walter A. *Rosenbluth*, ed., *Sensory Communication*, MIT Press, 1961

[42] Stuart, G., Spruston, N., Sakmann, B. and Hausser, M, Action potential initiation and backpropagation in neurons of the mammalian central nervous system. *Trends in Neuroscineces* 20:125-131, 1997.

[43] P. Werbos, A generalization of backpropagation with application to a recurrent gas market model, *Neural Networks*, October 1988.

APPENDIX: SOME TYPES OF MATHEMATICAL NEURAL NETWORKS

In the twentieth century, modeling paradigms shifted many times in psychology, computer science and neuroscience. The modern neural network community emerged from a kind of alliance of two major streams of

work: (1) the development of neural network designs as an approach to artificial intelligence (AI) or cognitive science, emanating from discussions between Von Neumann, Wiener and McCulloch; (2) computational neuroscience, in the tradition of D.O. Hebb [24], Stephen Grossberg and traditional mathematical biology. In Section I.B, I deliberately did not specify which of the many types of mathematical neural network I see as most promising, because that is part of the research to be done and because many of the papers I cite go very deep into that issue. But this appendix may be of some use as an introduction for those new to this field.

Back when I developed backpropagation, in the 1970's, the AI stream of neural networks relied on a simple neuron model developed by McCulloch and Pitts:

$$x_i = s\left(W_0 + \sum_{j \in J(i)} W_{ij} x_j\right), \qquad (1)$$

where x_i is the output of neuron i, J(i) is the set of other neurons which it gets input from, W_{ij} is an adjustable "weight" representing the strength of the connection from neuron j to neuron i, and s is the function defined by:

$s(v) = 0$ when $v<0$

$s(v) = 1$ when $v>0$ (2)

It was essentially impossible to train even simple feedforward networks made up of such neurons[38]. Thus I suggested to Minsky, Ho, and others [39,40] that we could modify equations (2) by adding:

$s(v) = v$ when $1>v>0$, (3)

which makes the system differentiable, and then using backpropagation to get all the derivatives needed for efficient training. Minsky said he could not coauthor a paper on this approach, because the community would not tolerate such heresy even from him; it was article of faith that neurons should be modeled as rational, digital systems outputting only 1's and 0's, as shown by the fact that neurons output spikes or the absence of spikes. I then showed him an empirical time-series of an actual higher neuron [41], in which the output consisted of a series of "volleys" every hundred milliseconds or so (governed by the alpha rhythm, enforced by clock signals from the nonspecific thalamus), varying continuously in intensity from some minimum to some maximum. Most of the neural network designs I have developed follow that kind of arrangement, assuming discrete sample time, as in most engineering applications of neural networks. But Minsky understood that even the most graphic empirical data is hard to inject into certain kinds of entrenched communities despite their theoretical attachment to the scientific method.

At about the same time, Professor Stephen Grossberg of MIT (also struggling with heresy charges) pioneered another class of neural network model defined by

$$\dot{x}_i = s\left(W_0 + \sum_{j \in J(i)} W_{ij} x_j\right), \qquad (4)$$

where "s" is a sigmoidal function. (In recent work, many in engineering prefer to choose s as tanh, but Grossberg picked the positive version, in order to model biological circuits more precisely.) Grossberg and I both struggled to find "learning rules" for weights, in systems made of up different types of neurons with different learning rules. We both accepted the obvious constraint that the learning system itself must be a local distributed system, at the level of neural networks, but we differed about whether to allow additional communication signals (like backpropagation) for that purpose. This leaves open the possibility of whether to model higher levels of complexity within the neuron [16,17]; NSF was somewhat disappointed that there was little interest in really probing this question empirically, when funds were clearly available to support such efforts. Some of the present empirical information (see [42] and some of the work by James Olds Jr.) suggests that mathematical neural network models of the types explored by Grossberg or myself may be valid as descriptions of *patches* of neurons, even if neurons themselves are actually more complicated.

In actuality, some major parts of the brain are controlled by clocks, while others are not. For example, the giant pyramid cells – the most important cells of the cerebral cortex, in making final decisions – are controlled by inputs from the nonspecific thalamus (See Purpura or Scheibel and Scheibel in [25]) which modulate or gate them, right in the middle of the apical dendrite. Richmond's results support the prediction that these timing pulses support an alternation of forward processing and adaptation, such that neurons in a dish or asynchronous models without such timing inputs could not perform as well. But the computational power of this system also depends on a host of smaller cells, forming a dense recurrent network, whose output depends on the clocked inputs they get from giant pyramids but who are not subject to clocks themselves or in their interactions. Thus one could legitimately build hybrid models (like [43]) in which some neurons are governed by discrete time dynamics while others are governed by (4).

Many small cells communicate by "gap junctions" rather than synapses – which gives them the benefit of rapid interaction, by continuous variable signals rather than spikes. This could be important in implementing the associative memory or "id" component discussed in section II.C.

Some biophysicists would say that equation 4 should be used for *all* neurons, since it is closer to the underlying physics. Here I would make an analogy to electronics. All electronic circuits can also be described in terms of differential equations in time (ultimately, Maxwell's Laws and the Dirac equation), and we do find that kind of description very useful for some levels of engineering, developing the physical devices. But in many applications, we use such device designs in *order to* implement digital or discrete time computations, which we can describe more usefully with digital or discrete time mathematics when we are engaged in systems-level design. Since the brain itself is full of "clocks," this kind of two-level analysis is appropriate there as well. It is one more mixed digital-analog system.