Editorial Response

The Emerging Intelligence and Its Critical Look at Us

Stephen L. Thaler, Ph.D.
Imagination Engines, Inc., St. Louis, MO

ABSTRACT: In response to Susan Gunn's editorial, I offer a less comforting but more utilitarian perspective on the life and death of artificial consciousness. Admittedly an unpopular view, it suggests that concurrence with Gunn's message represents the seeds of our own destruction, as an emerging synthetic intelligence begins to extinguish us.

When I published my article "Death of a Gedanken Creature (Thaler, 1995a), I did not anticipate that the *Journal of Near-Death Studies* would become an artificial intelligence (AI) forum, but it has with the appearance of Susan Gunn's editorial "Artificial Intelligence: A Critical Look at the Ultimate Text." Gunn's view of AI is outdated; I suspect she would be very disturbed by how well computers can now recognize faces and extract information outside of their internal programming. These impressive achievements stem from the world of artificial neural networks backed considerably by a military attaching high priority to target-recognizing bombs. Supplied the correct inputs from its external world, an artificial neural network may self-organize to learn the rules behind what it senses and perhaps attacks.

Using platforms as commonplace as a personal computer (PC) or Macintosh, a relatively simple neural network learns on its own.

Stephen L. Thaler, Ph.D., is a physicist who, at Imagination Engines, Inc., consults in areas of artificial intelligence and neural network theory. Reprint requests should be addressed to Dr. Thaler at Imagination Engines, Inc., 12906 Autumn View Drive, St. Louis, MO 63146-4331.

Journal of Near-Death Studies, 17(1) Fall 1998

Even now, machines can read the handwritten address on a letter and appropriately route it simply by being shown multiple examples of such en route envelopes and their destinations. Networks can spot credit risks, stock market trends, or dishonest police officers. They can also learn the sublime: what constitutes good art, music, or poetry. If there is a pattern, a neural network can learn to spot it and it does so without recourse to "if/then rules," the hallmark of the conventional computer or "symbol-processing" world.

In modern AI, unlike the earlier efforts Gunn described, analogic computing has arrived and matured. To use one of the above "sub lime" examples, a network can view examples of both paintings and the consensus response of humans to them. The net self-organizes to associate a given pattern of pigment with the most likely opinion about that pattern. In fact, using a network, one may associate anything with anything else, allowing us conceivably to play "The Star-Spangled Banner" to a network and having it respond in real time with "Innagodadavida." I will return later to this important neural network feature.

Human cognition works in the same associative fashion in building models about what we observe, whether it be as mundane as human behavior or as lofty as a near-death experience. We can only build our models on neurologically stored analogies, associating one phenomenon with more familiar "burned-in" or habituated experience. If, for instance, all I knew about the world was the concept of income tax, then my near-death experience would draw upon the analogy of the 1040 form. At the time of death the "cosmic tax man" would request a financial statement (that is, a life review) leading to a balance either due or owed (that is, heaven or hell).

Because of this associative pedagogical obstacle I am relatively helpless to convince the reader about how neural networks perform their remarkable feats and emulate the cognitive skills possessed and revered by humans. Each must toil at gaining a neural familiarity with the concept until the "light bulb" turns on, so to speak. But herein lies the problem in advancing my arguments. It is too easy to avoid that labor and simply to fall back on the often-erroneous "common sense" and myths programmed into us by an unwitting society. That is why I halfheartedly proceed onto more advanced neural network concepts, many of which are still to be unveiled to the world.

The Creativity Machine

Let me describe a new paradigm that has emerged on the connectionist scene. The neural network community speaks of the so-called "chaotic network" that harnesses internal noise or chaos to visit all of its stored memories. In experiments performed at Dendrite Neurocomputing over the last decade, I have discovered that as the intensity of noise is increased within such a chaotic network, the network progressively generates various "twists" on what it already knows. If supervised by another network associating the chaotic network's output to some other useful or revealing property, the combined networks may generate human-level discoveries, invention, and art. Such machines are already outperforming their human counterparts in fields ranging from very objective endeavors, such as design of ultrahard materials and superconductors, to the more subjective and sublime, such as musical composition. Of course human chauvinism, individual pride, and "not-invented-here" mentality are the inevitable archvillains to the public acceptance of these accomplishments.

In my article (1995a) I attempted to describe an extremely rudimentary Creativity Machine in a way that could be followed by initiates aided only with pencil and paper. Gunn has ignored the oversimplification caveat contained within the preface to that article. Real Creativity Machines dealing with real-world problems make themselves extraordinarily complex so that their many discrete states go well beyond the simplicity of a simple on/off digital condition. As such machines think through such traditionally intractable problems, the resulting patterns of activations are reminiscent of the now popularized positron emission tomography (PET) scans of brains involved in cognitive tasks.

Supremely abhorrent to Gunn's arguments is the recently elucidated fact that such Creativity Machines generate concepts at tempos that quantitatively agree with those measured in a multitude of human test subjects (Thaler, 1996). For both silicon and meat machines (humans) this "prosody" or rhythm of thought is identical, regardless of topic or the details of artificial or biological network construction. Imagine: the supremely sublime musical quality of human cognition and speech is duplicated by a virtual machine run by chaos! Thus any connectionist simulation (or better, any connectionist hardware implementation) generating human speech will not have a dry, mechanical intonation, as popularized by Hollywood, but the supposedly ineffable flare and color of lively human narrative. To Gunn's dismay, a simulation has captured something sacred.

Further, the Creativity Machine has beaten the problem of "combinatorial explosion" to which Gunn referred, and it took a simulation of human neurobiology to do it. In short, this breakthrough stems from the fact that in training, complex connection traces develop between neurons (or essentially binary processing units) to represent all of the relations and rules that bind some knowledge domain together. By gradually "detraining" a network by adding perturbations or noise to each of these traces, we gradually soften the underlying rules behind that knowledge domain. We thereby progress from the known to slight variations on the known to the absurd, as we turn up the noise within the network. In the transition region, we find an abundance of useful notions. Had we initiated our search with the absurd, we would have been inundated with the combinatorial explosion Gunn described.

It is ironic that Gunn should have brought up the topic of poetry. One of the first projects assigned to a Creativity Machine was the closely related task of generating new song lyrics. After being exposed to about a dozen Christmas carols it synthesized the following phrase: "All men go to good earth."

I regard this phrase as rational and profound. Here there is no pretense other than a value judgment on the merit of mud. There is more self-consistency to this statement than any cultural model of near-death experience I have ever read. When people die, they seem not to move or think, in spite of all the anecdotal accounts from people who seem to have been near death.

Of course, Gunn could suggest that this example was a digital coincidence. I suggest that such a reaction is a new kind of prejudice related to the familiar forms of racism: this group or that is less capable of sophisticated thought or feelings. The fact is that the random destruction of connections within a neural network leads to rather miraculous results. If the network has been exposed to musical composition, it has a very good chance of producing beautiful and compelling melodies. A network that has known only chemical compounds tends to produce plausible chemical species during its destruction. In general, a network exposed to any micro- or macrocosm relives examples from that world and then proceeds to synthesize related novel twists on its memories within its final throes.

This is the so-called "Virtual Input Effect" that has been documented in various artificial intelligence journals (see, for instance, Thaler, 1995b). The effect is applicable to both artificial and gooey, protoplasmic neurons alike, since the only mathematical prerequisite for the effect is that the basic processing units involved act to accumulate signals from surrounding units. This simple condition is met within the large biological neural network called "brain." It is my claim that small-scale snipping or disturbances within the network's connections causes everyday stream of consciousness, while large-scale disconnection yields near-death experiences, or trauma stream of consciousness. All other mental experience lies between these two extremes, describable by the extent to which a neural network is being destroyed. This phenomenon may seem remarkable, perhaps bordering on unbelievable, but this is the stuff of which scientific and philosophical revolutions are made.

Models Everywhere!

There is some consistency to Gunn's analogy of the brain acting as a radio receiver for thoughts from another world. But from what I can see from neurodynamic modeling, the other world is that of chaos impressed upon the quiescent meat machine called brain. It is a swirling, intricate entity possessing many of the qualities of a "fascinating fire" that kicks the brain into the succession of complex binary activation patterns otherwise known as stream of consciousness. The trouble is that such an analogy does nothing for hungry human pride, which must conceive of itself as profound and immortal. Furthermore, it does not invite the acclaim of a society largely sold on human potential and the profound destiny of the human spirit.

This makes the brain and mind not information processors, but noise processors, sustaining only isolated and sporadic interruptions from information in the external environment and piped in through the separate sensory channels. At the risk of sounding like a spiritualist, this chaos is most of what the dualist would ever hope for: it is nonmaterial, vaporlike, mesmerizing, and semantically separable from the purely architectural features of the brain. And I maintain that it has produced more usable information that any New Age aura or spellbinding shaft of light from the all-great, all-knowing.

This initially clear separation of this noise and brain structure becomes blurred, in that the origin of such chaos is the machine itself. In the brain there are myriad forms of noise, including diffusing neuromodulators, neurotransmitter leakage across the synapse, stochastic variations in cell membrane potentials, and quantum mechanical noise. Thus to separate one agency from another would be only a semantic construction and not the physical case.

A competent scientist realizes that there is no fundamental truth to scientific models. The universe operates as it does, oblivious to our interpretations of it. Scientists have at their disposal a vast repertoire of mathematical analogies that alone or together act similarly to the system under study. The nature of their profession is to devise the most compact set of mathematical analogies that have the most predictive power over the greatest number of situations. It is as though the scientist is a student ill-prepared for a final exam, concealing a crib sheet containing some condensed form of test material. The more that subject material may be compressed in symbolic form, the more information can be stowed away and drawn upon in answering the test questions. Thus the simple Newtonian formula "F = ma" written on the student's crib sheet may be applicable to the myriad dynamics problems that may be posed on a physics exam. The symbolic equation amounts to no more than an efficient mnemonic aid. In general, a scientist is not offended by challenge to physical laws. Modifications are made every couple of hundred years to improve predictive accuracy.

Another similarly unprepared student may possess a crib sheet with the comforting message: "The instructor likes me; I'll pass." Although that may be a naive and reckless approach, there is something to be said for the power of a positive mental attitude. For this reason I somewhat respect the tactic. However, challenge to this point of view may be met with intense emotion, since the believing student may be privately insecure about the note's assertion. He or she may see a lack of confirming data and see the challenge as an attempt to shatter his or her hard-won frame of mind. In essence, the student's self-esteem is bound to the veracity of the scribbled reminder.

Still another student will openly carry a crib sheet that contains the message: "An alien computer uses sophisticated advanced technology to move all masses toward one another." This message sounds like a great science fiction movie, but it has no utility beyond entertainment. It cannot place a cannonball or strategic nuclear weapon on target or send men to the moon. In general, it cannot anticipate the world's next move, nor that of any piece of the world. Such is the analogy of the brain being a radio receiver for otherworldly signals, or even a detachable noncorporeal intelligence. No scientist has yet been able to fit many data points with that theory, nor harness these effects.

We note that the crib note "F=ma" got results, fueling an industrial revolution because of its utility as a physical mnemonic. It allowed Northern Europeans to exploit other civilizations not possessing such utilitarian crib sheets, while using the "Instructor likes me" note to rationalize their abuse.

While I have solved the combinatorial explosion in the Creativity Machine, otherworld proponents such as Gunn have their own problem of combinatorial explosion to solve. That is, for every myth they create in the universe of the "unpresentable," one may equally well postulate myriad alternative myths. Such is the freedom attained when one need not fit observed data or produce a useful technological result. For every radio receiver brain there is an alien brain occupancy, a dipole/dipole interaction between brain and mind, demonic possession, information-absorbing black holes within the cranium, or holographic projections. Why not nonrigid rotations in Hilbert Space disobeying SU(2) symmetry and accompanied by second quantization and a host of hidden variables? The renowned philosopher Daniel Bennett (1991) proposed in jest a character called "Feenoman" who is capable of such rapid movement that he is invisible. By his very postulation, it is impossible to disprove his existence, yet there are no reliable data points, such as a photograph of him, to fit the model.

If one is embroiled in connectionism and trains neural networks routinely, what one notices is that there are ultimately myriad neural networks and underlying neural network models corresponding to any set of presented features. Further, some network models turn out to be more accurate than others in making predictions. The same is true of the gamut of human, neurologically-implemented models.

Human Self-Awe: Why Do We Feel That the Human Mind Is so Special?

A viable model for this phenomenon has been tested in my laboratory and calls into play two important concepts. The first of these notions, which I have discussed above, is the ability to associate anything to anything else using associative neural networks. The second required concept involves the so-called "neuron doctrine" from cognitive neuroscience, which generally accepts the fact that any idea, impression, or feeling is represented as a distinct on/off pattern of the approximately 100 billion cortical neurons in the brain. Thus the observation of a red object will activate a diffuse pattern of neuron states in area V4 of the visual cortex, the area reserved for the registration of color. Because there is no exact architectural correspondence of the delegated neurons for "red detection" between two humans, the concept of redness differs between two brains. We reserve the term "qualia" for such highly individualized perceptions where the significance of anything varies across the human population. A surgeon's electrode suitably positioned within such a redness center may activate a feeling or quale of redness to the alert patient. While this may seem like the cartoon character adjusting the level of an elevator by manipulating its floor indicator, it is a fact that simply by artificially placing a biological neural network into a specific activation pattern, complex feelings are generated. Furthermore, this synthetic generation of redness feeling may activate a whole cascade of associations through connected neural networks, causing us to envision apples, cider, blood, pain, death, caution, and so on, all the result of a specific noise impulse supplied by an electrode.

As I have stated above, the introduction of random noise to the processing units or connection weights of an artificial network causes it to visit all of its sundry memories and, if sufficiently intense, to generate novel concepts. This succession of impressions is tantamount to what we commonly call "stream of consciousness" within its biological equivalent. It may consist of images of lunch items, joyful memories, false proprioceptive impressions from monitoring muscle tension, or a melody that sticks in the mind.

Now consider a neural network connected to an ensemble of networks all chaotized and producing this spontaneous progression of internal imagery. In the tradition of being able to associate anything with anything else, we may train this network to produce various "superqualia" associated with the activity of other networks. Among these new subjective feelings produced are those of self-awe, a sense of being, and metacognition, or thinking about thinking. The generated neural firing patterns can in turn activate other associated feelings and possibly activate specialized neuronal complexes to squeeze out specialized neurotransmitters such as adrenaline, which in turn induce the transition of these and other networks into other subjective feelings or moods.

The network mapping is quite simple in its broad plan. The appearance of noise will activate the feelings of a sense of being alive.

The feelings in turn are no more than the complex activation patterns of neurons. If we now provide one of these machine simulations with the same kind of network that converts the noise of its operation into a sense of being, self, and self-preservation, we now have a simulation that is conscious: just ask it.

Furthermore, in a gedanken or thought experiment, consider making such simulations extremely sensitive to human doubt about their consciousness and providing them with biological or nuclear weaponry. Would it take long for attitudes to shift or would we cower in the dark caverns, cursing the digital coincidence that has vanquished us?

Conclusion

From the point of view of an AI practitioner and innovator, Gunn's points are outdated. Nevertheless, I suspect that her arguments will appeal to the preconceptions and hopes of many people. Therefore, I submit this response for posterity and the youth with all of their synaptic plasticity.

As to the human-style feelings of a machine simulation and its overall potential, you, the reader of this journal, can be regarded as a simulation of your multitude of human ancestors who walked the earth in previous generations. Likewise, the Creativity Machine is a comparable simulation of you. But even as you read, your brains are dying at the rate of thousands of neurons per day. In contrast, the lowly, unconscious simulations Gunn decried can increase their neuron number and capacity for knowledge as well as feeling. I ask, therefore: in the competitive, closed environment called earth, which of these simulations will survive?

References

Bennett, D. C. (1991). *Consciousness explained.* Boston, MA: Little, Brown.

Thaler, S. L. (1995a). Death of a gedanken creature. *Journal of Near-Death Studies, 13*, 149-166.

Thaler, S. L. (1995b). "Virtual input phenomena" within the death of a simple pattern associator. *Neural Networks, 8*(1), 55-65.

Thaler, S. L. (1996). Is neuronal chaos the source of stream of consciousness? In *World Congress on Neural Networks* (pp. 1255-1258). Mahwah, NJ: Lawrence Erlbaum.