WHAT NEURAL NETS STILL CAN'T DO

Foundations of Cognitive Science, Prof. Peter Slezak

Okko Buss, 3152903

June 2005

The developments of computational theories and computing technologies in the latter half of the twentieth century have had far reaching effects both on the debates among philosophers of mind and the range of tools and concepts available to computer scientist. Engineering endeavors in Artificial Intelligence tend to fall back on and both, providing a testing ground for the tools of the latter in application on the former.

In the 1980s, developments in Neuroscience had a similar effect as computing had on philosophy and engineering earlier. Questions of how the brain (i.e. the "brain-mind") might represent and compute seemed explicable in terms of neurology. Parallel and distributed computer models entered the AI toolkit, with some impressive initial success. This approach is particularly associated with Paul and Patricia Churchland, who continue to forge a unified philosophical, scientific and engineering approach to Cognitive Science.

While symbolic AI, both weak and strong, is seeing a comeback recently (commercial applications of speech/language technologies and expert systems are becoming increasingly mature, CYC is scheduled to go online during 2005), the addition of neurology and neural nets to the philosophical debates and AI toolkit has proven fertile. The arguments that the Churchlands have offered for their case, and their rebuke of common criticism from computationalists are compelling, even though their reductionist flavor has many skeptics. This paper will present and flesh out Churchland reductionist connectionism. It will further critically evaluate their arguments, drawing on criticism both from within and outside the symbolic AI and computational functionalist camps and conclude with some remarks on the possibility of breaking the deadlock between the two sides.

Neural nets have indeed had great effect on understanding how the brain might represent and compute. Two early successes of modeling cognitive tasks on a neural net , text to phoneme generation (NETtalk) and edge detection, are described elsewhere in greater length (Churchland & Sejnowski, 1989). In brief, what follows is a recapitulation of how phoneme generation from input text and edge detection from shaded images works on a neural net, and what the significance is for representation and computation.

Describe NET-talk TTS

Describe edge-detectors from shaded images.

These two serve as examples of innately human capacities that have been successfully modeled on neural nets, which has also offered some explanation as to how these capacities might function in humans. This is in stark contrast with most advances in classical/symbolic AI, which have been largely concerned with emulating human intelligence. A well-trained chess machine does not help us learn anything about how humans play chess.

Another, more recent example that demonstrates the prowess of neural nets in modeling cognitive tasks closely to their human implementation, is the work of Marc Leman (1995) on tone center (harmony) recognition. Recognition of harmonies is an isolated cognitive task, and as such it offers an interesting example of how innate capacities might be built and represented on a neurological basis. Some parallels between language- and music-specific tasks might help to illustrate why tonal recognition is such a capacity.

Tonality in music works much like syntax in language. Its acquisition is a

natural, unguided process of human developmental psychology. Everybody with a healthy hearing has a clear but intuitive understanding of what is consonance and dissonance. That intuition varies across cultures, who have a different common understanding of tonality. There may even a poverty of the stimulus argument to be made for acquisition of tonality occurring in spite of broken and incomplete input. Lastly, tonality is a context-sensitive system, in which constituent chords and harmonies have meaning in virtue of their structural relation to other chords and harmonies.

The relative isolation of tonal recognition from other tasks in music cognition, such as recognizing timbre of specific instruments or melodies of a particular tune (e.g. we can recognize chords regardless of what produces them, or what piece of music they occur in), allows for treating tone center recognition as a n autonomous task (similar to text-to-phoneme conversion or edge-detection). This isn't straight-forward in other innateness-candidates, such as syntax. Thus, simulating tone recognition on a neural net should prove valuable to how innate human capacities are implemented and represented in brains (or brain like structures).

Leman's experiments, like the analysis of NETtalk and edge-detection neural nets, showed that after sufficient training a neural net could accurately perform a characteristically human task. He trained a neural net with 300 iterations of Shepard triads (particularly clean chords) from the western musical tradition. Analysis of the neural net was performed after one, 60, 150 and 300 iterations. The method analysis was done on so-called Kohonen maps, which allow graphical representation of a neural net according to Euclidian neuronal distance, based on connection weights between neurons.

By this method it was possible to identify a characteristic output neuron for each chord/harmony presented to the neural net and its distance to other such neurons (associated with different harmonies). Common musical patters were identifiable by characteristic patterns on the map, mirroring systematicity of music, for which humans have strong intuitions.

Most researchers who employ neural nets to model cognitive tasks agree on some of the limitations they come with. These limitations are usually about scope (how specific/general can the to-be-performed task be) and the analogy between artificial and real-brain processes (especially backward propagation as a learning algorithm and random starting configuration of artificial neural nets are questionable). However all are enthusiastic about the philosophical and technological possibilities that experiments such as the three discussed above have offered. So where is the criticism?

While there is micro-level criticism even among connectionists about the limitations of artificial neural nets as real-brain models, the real stand-off occurs between the connectionist and two opposing camps: Symbolic AI engineers on the one side and computational philosophers on the other. There are three general criticism to connectionist AI on offer: "neuroscience is too difficult," psychological states have "multiple instantiability" and "intentionality (Churchland, 1989, page 19f)." Before evaluating the connectionist response to each of these criticisms, let's have a look at each in detail.

The claim that neuroscience is too difficult appears slightly out of character on that list of three. Understandably, this seems to be the one most

briefly addressed by the proponents of connectionism. It is hard to imagine why the difficulty of a task should be a deterrent for conceptual reasons. However, as I shall explore later, the criticism is also one of the inherent trade-off of effort over explanatory prowess. In other words, how much "bang for the buck" does connectionism offer.

The multiple instantiability argument comes largely from computational functionalists such as Hilary Putnam, Jerry Fodor and Zenon Pylyshyn, but is relevant to the role symbolic AI plays in cognitive science. If cognitive processes can only be purely in the language of neuroscience, the explanatory power of the entire computational AI toolkit is at stake. The idea behind this criticism is that if a psychological state can be instantiated in more than one way, an attempt to describe it only in terms of one will be insufficient. Instead, description of psychological states in terms of high-level observations (often from folk psychology) avoids unnecessary commitment to questions of implementation.

Lastly, the intentionality criticism seems to come from several philosophical camps, computational functionalism and John Searle alike. Intentionality of psychological states, such as beliefs and desires, is a result of such states being "about" other things, which they represent, but to which they are not directly causally related. What's important then is the semantic relation that mental representations hold to their targets. Because of this special type of relation, neurology cannot explain intentionality because it works only on explaining causal relation of the "brain-mind" to the world.

This view is most pronounced in Zenon Pylyshyn's "classical architecture" of computation. Within a computational system, there exist a

"semantic level" (what the system knows), a "symbol level" (encoded knowledge) and a "physical level" (implementation). The three are kept strictly separate, so that the workings of one level cannot be explained in terms of the other two. For the connectionism debate in particular, intentionality is concerned with the semantic level, which cannot be explained in terms of the physical, implementation level.

Connectionists/reductionists (in particular Patricia and Paul Churchland) have responded to these accusations. The general reply has been to point out that critics of connectionism seem to tolerate a conceptual inconsistency, which they term "Theory Dualism (Churchland, 1989, page 19)". The inconsistency in 'Theory Dualism' is a result of believing on the one hand that mental events are not a explicable in terms of Cartesian dualism (i.e. body and soul), while never-the-less denying that mental states can be explained in terms of pure neurology on the other. Patricia Churchland, with Terrence Sejnowski (creator of NETtalk) have outlined this characteristic of anti-reductionists and forged a response to the above three 'Theory Dualist' criticisms in detail.

The first criticism, about the difficulty of neurology, they claim is hardly an objection to pursuing it at all. Difficulty is not a *conceptual* objection and there's every reason to believe that states of the brain and psychological states are systematically related. Progress in technology and techniques will remedy the inherent difficulty of neuroscience in due time, the response concludes. This response is kept brief.

Although this response addresses the fear that neuroscience is a difficult enterprise by pointing out that, like all science, it will become easier over time,

it doesn't address the question of whether it is a *worthwhile* enterprise that stands behind this fear. The inherent trade-off in explanatory power and effort of explanation that underlies this criticism is largely glossed over in the reductionist answer. As we have seen, easily isolated cognitive functions (such as text-to-phoneme conversion, edge-detection or tone-center recognition) might be easily identified and modeled. However other psychological phenomena, such as schematic knowledge or scripts (e.g. what do we know implicitly when we sit, order and eat at a restaurant and how is that knowledge represented?) may very well be beyond neurology and connectionist AI.

In that light, the question becomes one not only of whether it is feasible but whether it is worthwhile to invest in difficult explanation of simple behavior. However since folk psychology does not offer a satisfactory way out of this dilemma, this is largely glossed over by reductionist optimism. I will revisit this claim when discussing issues of control in intelligence.

The second criticism of multiple instantiability is given an equally brief, but somewhat more satisfying answer. Even if high-level mental states can have a variety of low-level implementation, it does not undermine the claim that systematic correspondence between neuronal and functional activity in the brain means "a reduction suitable to that domain (Churchland, 1989, page 21)", which is sufficient and suitable an explanation. Two cars running on methane and gasoline don't prompt a conceptual change to our understanding of the internal combustion engine.

The third, intentionality criticism, poses the greatest point of friction between reductionist and computational functionalist camps. This is largely

because the explanatory relevance of two central notions of Cognitive Science is at stake: computation and representation. The connectionist response to the claim that neuroscience cannot account for intentionality draws on what connectionists see as severe defects in anti-reductionists taking these two, computation and representation, at face value.

In symbolic AI and its philosophical counterpart, computational functionalism, computation in Turing's, i.e. the structural, formal sense is a notion central to explaining cognition. This formal sense of computation involves the manipulation of meaningful symbols. Intentionality is then explained as the computational manipulation of meaningful of symbols, usually sentential attitudes from folk psychology (e.g. Language of Thought). For most of its history, discrete and digital computers stood as sufficient testing ground for the computational theory of the mind. This is where the connectionist response to the intentionality argument picks up.

The idea that cognition is computation in the above sense is flawed since it is too dependent on the serial computer metaphor (or model). There are a host of reasons for why the comparison between brains and digital computers (Turing Machines, von Neumann architectures) cannot account for real human cognitive processes. Patricia Churchland cites the following (Churchland, 1989, page 25):

Human physiology too slow for sentential processing (therefore there cannot be a Language of Thought underlying *all* cognition)

The brain is not a serial computer (some tasks are more effortful in one than the other, others less effortful)

The brain's memory is content addressable (not location addressable)

The nervous system can degrade gracefully (not 'brittle', like serial computers)

Hardware-software distinction is misleading (they are the same in neural computation)

These reasons demonstrate that taking the mind-computer metaphor at face value is a risky enterprise, if not fundamentally flawed. Whatever role computation plays in cognition, the idea that cognition is meaningful symbol manipulation and that beliefs and desires are sentential attitudes would defy most of the things that we know about human physiology, which eventually should to be consulted for how cognition is implemented, claim the reductionists.

One last argument offered against the intentionality criticism (this one less concerned with the nature of digital computers) is that non-verbal animals and infra-verbal humans pose a major problem for the claim that beliefs and desires are sentential attitudes in the Language of Thought. Obviously, in these instances cognition must be accounted for without falling back on a notion of sentential attitudes.

These two sides of the debate over the relevance of neuroscience and artificial neural nets to Cognitive Science seem to be in a deadlock. However initial successes of connectionism in AI and the failure of its critics to find steadfast conceptual reasons to undermine its claim that neurology must be a key notion to understanding cognition have essentially secured its place in Cognitive Science canon and methodology. What's left unanswered is the question to which extend connectionism can successfully be reductionism.

In *The Minds New Science*, Howard Gardner answers this question cautiously. Neurology, he claims should represent a "lower bound (page 287)", much like cognitive anthropology represents a natural upper one. Although overall he strikes a more reconciliatory tone than most critics, his reason for caution is that neuroscience, if unguided, is wont to 'miss the point':

> "[T]o take an example from language, a neurologist ignorant of linguistics might rely on naive intuitions about language: one would therefore describe an aphasic patient as unable to use "small words"... a linguistically trained observer will immediately be able to to pose questions and introduce distinctions at a subtler level. (page 287)"

Neuroscience without Linguistics, Philosophy and Psychology would indeed be a blind undertaking.

A further, more fleshed-out criticism of reductionism comes from Aaron Sloman at the University of Birmingham. A philosopher turned computer scientist, Sloman describes intelligent systems as architectures that govern mechanisms and substates. The intelligence of the overall system is a result of a minutely tuned interplay of its various components. This is said to be true of living and artificial intelligence alike.

Before outlining his criticism of reductionism, if may be worth noting that Sloman also heavily criticizes pure computational approaches. He claims that computation explaining intelligence falls short on several counts. Firstly, computation is a formal, structural notion and therefore cannot account for causality and control in intelligent systems. (Extending computation to include causality leads to trivializing the concept or creating circular definition of intelligence.) Secondly, there are essential processes in intelligent agents that cannot be accounted for in terms of computation at all (e.g. chemical ones). He therefore calls for an open mind when determining what processes

are required for intelligence. Cognitive Science must create a "taxonomy" of intelligence before truly understanding and modeling it.

However his reproach of computation as a key notion for explaining intelligence does not make him a fan of reductionism either. A "bottom-layer" approach to causality in intelligent architectures cannot exist, he claims. To demonstrate the absurdity of this idea he falls back on the example control in a computer. A program does not have full control of the processes it runs, since it has to share machine resources and is in turn controlled by an operating system. Furthermore, a high-level program (or code in a high-level programming language) has little or no insight into the processes it spawns (or the program after compilation.) Low-level machine code and even lowest-level voltages also cannot be said to 'control' the computer since they just follow instruction from high-level processes, or as sit as passive code elements. Control then is a question of distribution among more or less 'virtual' processes in the overall architecture.

The concept of a virtual machine, a high level process implemented on a lower-level architecture, aptly shows why an architecture-driven approach is useful for explaining how control is distributed among components within an intelligent agent. A high-level process (e.g. a word-processing program) still has control within the overall system, as do the bit patterns in the of the CPU. The question is not 'who or what controls the system?' rather than 'what level of control do the components have within the overall architectures?'.

If reductionism fails even to account for control in well-understood architectures like modern computers, failure in explaining human beings seems like the obvious conclusion. Sloman sums up:

"Some readers may be inclined to argue that there is only one

> 'ultimate' level of reality and only at that level can processes be controlled. This requires drastic rejection of the most common-sense concepts of causation and control. (Sloman page 188)"

Two things may be worth reprising and evaluating before concluding the discussion of Sloman's view on computation and reductionism. The first is that, he shares with the reductionists the view that computation is insufficient for explaining intelligence, even if his aim is not to get rid of it. In fact, the idea behind intelligent architectures seems to be more about finding a proper place for computation in the range of possible mechanisms that comprise intelligence.

The second point is that his idea of 'control' stands is particular in contrast with Pylyshyn's three-levels view of computation. In Sloman's view, control/causality is shared by processes on various layers in a complex system (where the directness of control is inversely proportional to the level of abstraction of the process, the higher, the less direct the control). In Pylyshyn's view, high-level processes cannot even be considered alongside or in terms of lower-level (i.e. "physical level") ones.

In this regard, he stands somewhere in between the fronts of reductionism and computational functionalism. He shares with the former a profound doubt that purely computational approaches will be able to account for the entire range of possible psychological phenomena even if he wants to rehabilitate computation. And with the latter he shares disbelief that a pure bottom-layer approach will be fruitful, even if he does not grant any other level of explanation such powers. In accomplishing this (neither giving up the notion of computation altogether and nor denying the relevance of low-level implementations of higher-level processes) his taxonomy of intelligence stands a good chance of opening new avenues of investigation for a debate

long locked in stalemate.

In the end, this warrants a reinvestigation of the difficulty criticism, not because difficulty is a conceptually irrelevant, but because difficulty can be avoided by choosing cheaper, equally valid answers. Sloman proposes an approach that allows for high-level interpretation in terms of virtual machines in intelligent architectures, without falling back on faulty assumptions of folk psychology with which the reductionist take issue. In light of this possibility, I suspect that there are explanations that the high-level computational approach has to offer that can be salvaged. Explanations, which in terms of connectionism, even if possible and valid, would be far to strenuous and difficult.

REFERENCES

Sloman, Aaron. "Beyond Turing Equivalence"

Leman, Marc. "Music and Schema Theory. Cognitive Foundations of Systematic Musicology"

Pylyshyn, Zenon.

Gardner, Howard. "The Mind's New Science"

Churchland, Patricia & Sejnowski, Terrence.