Graham Morehead
Language and Cognition November 22, 2004

Abstract

Could the human brain ever understand itself? Several definitions of the parameters involved will be discussed. What does it mean to "understand"? What are some implications of self analysis? Arguments will be presented from Gödel, Wittgenstein, and Chomsky. An attempt will also be made to describe a framework for autocomprehension.

Humans and Rats share 90% of their genes (Lindblad-Toh, p475). How different can we really be? How different can our brains really be? Can the differences be *qualitative*, or are they merely *quantitative*? We certainly do not expect a rat to be able to solve any maze (Chomsky, p223). Could we expect a human to be able to understand any concept? What does it mean to *understand*? In this paper I will look at some questions and answers from past thinkers and afterward attempt to offer a framework upon which to construct some new questions.

Can a snake eat its own tail? Can a vacuum cleaner suck itself up? There is a well-known set of Unix software tools commonly referred to as "GNU", which stands for "GNU's Not Unix." What does GNU's really stand for? This and other self-referencing acronyms can be expanded to infinity without result. If the word *heterological* means "Does not describe itself", is *heterological* itself heterological? (Hofstadter, p21) The concept of self-reference has sparked many philosophical debates and much common interest - from Narcissus with his self-love, to modern self-help books, which are written and sold on the premise that we can indeed understand ourselves. Even the Holy Bible mentions the paradox given us by the famous Cretan, Epimenides: "All Cretans are liars" (Titus 1:12). How do we ask these questions in scientific terms? How do we answer them?

Mathematicians in the early part of last century were struggling to deal with some aspects of set theory. One such issue involved classifying all sets as either *non-self-swallowing*, such as "The set of all red vegetables", or *self-swallowing*, such as "The set

of all things that are not red vegetables". This second set is not a red vegetable, clearly. It ***is*** therefore included in itself. How would you classify "The set of all *non-self-swallowing* sets"? If you say it is itself *non-self-swallowing*, it becomes a part of its own set and must therefore be *self-swallowing* (Hofstadter, p20). The Epimenides paradox rears its head again.

Russell and Whitehead wanted to describe a formal system without logical quandaries like the one above. Building on George Cantor's work in the 1880's they detailed Principia Mathematica (PM). The flow of set theory complications were stanched in PM by limiting the scope of sets. Each set in PM belongs to a *type*. The lowest type of set contained only objects, not other sets. The second-lowest type of set could contain only objects and/or sets of the lowest type. In this scenario a set cannot ever belong to itself. It could only belong to a set of a higher type. Russell and Whitehead were successful in creating a formal number theory which did not include logical "strange-loops", as far as they could tell (Hofstadter, p21).

In 1931 Kurt Gödel published his most famous paper, "On Formally undecidable propositions of Principia Mathematica and Related Systems". He made quite a stir. Some even questioned the validity of known mathematics. Proposition VI states (basically) that any consistent set of formulas contains unprovable formulas (Gödel, p57). In a rigorous proof that is still admired today, Gödel proved that even PM must contain at least one logical strange-loop (a paradox like the ones described above) , without even needing to state it explicitly. There is no formal system that is able to prove every statement that can be made within its own framework; This concept is known as *Incompleteness*. A *formal system* is any system where the axioms are known and all transformational rules are known. Elements of a formal system can only be generated by operating the transformational rules upon the set of axioms.

How could the inner workings of the mind be described? As a formal system? A subset of cognition must be chosen - one that allows detailed inspection. Sleuthing out the theoretical transformational rules and axioms will require a large reliable data set. Language is a unique window on the mind's workings - both detailed and sharing similarities with other mental processes. Pylyshyn claimed that all thought is at least quasi-linguistic (Haugeland, p265). Present theories still leave room for an ambiguity

that comes from our own lack of knowledge, but the thoughts themselves are not so ambiguous. The brain thinks in *Mentalese*, as Pinker terms it. Nothing can be left to the imagination, because mentalese *is* the imagination (Pinker, 70). Thoughts must be represented in some format, it might as well be called *mentalese*.

How can an element of mentalese be studied? Many of the words used in this field suffer from the inspecificity of informal usage. What does it mean to understand? to know? to cognize? What is the meaning of a word? Ludwig Wittgenstein asked this same question. His first answer to this and other basic questions was that one should look at how one measures meaning. When answering the question, "What is length?", for instance, he ventures that one should first answer "How does one measure length?" (Wittgenstein, p1). How do we measure what a word means to someone? Wittgenstein holds that the meaning of a phrase is "characterized by the use we make of it" (Wittgenstein, p65). He goes on to describe a situation where a chess player puts a paper crown on his king saying that it has some meaning to him within the realm of the game, but does not alter the use of that piece. Wittgenstein says that the crown therefore has no meaning within the realm of the game. According to him, a linguistic concept has no meaning unless its meaning can be determined externally. The *game* in our discussion, is the entire world we live in. ***The rules*** determine how we process perception and ultimately respond to the world.

On the concept of *rule-following*, Chomsky says, "As persons, we attribute rule following to Jones on the basis of what we observe about him" (Chomsky, p242). John Searle disagrees with this idea by saying, "It is not enough to get rules that have the right predictive powers" (Chomsky, p247). Searle is most famous for his take on the *Chinese Room*. If mono-lingual English speaker is inside a chamber with a big book of Chinese phrases and proper responses he can pretend to be Chinese. When a question comes through the *in* slot he merely matches it graphically to the proper line in his big book and then copies down the proper response - another graphic. Although he has zero comprehension, the Chinese person on the outside may be fooled if the book is sufficiently complex and correct. Searle says we must have more ostensive evidence showing that an agent really is following a given rule. Chomsky explains in the following pages that it is the natural course in any science to fit a theory to available

evidence. When multiple theories match all available data, the simplest one should be chosen. He does, however, take issue with Wittgenstein on the point that rules only exist in the interactions between "players of the game", as it were. Chomsky says that rules are just as valid for Robinson Crusoe as for anyone because he is a person, like us (Chomsky, p233).

A concept from Artificial Intelligence (AI), called the "Ascription Schema" says that one can ascribe beliefs, goals, and faculties so as to maximize a system's overall manifest competence (Haugeland, p214). This would mean that you can ascribe beliefs and goals to a system based purely on that which is externally observable. Haugeland goes on to say that ascription is merely a way of discerning order that would otherwise go indescribable. Simple systems such as a mousetrap certainly have function, but ascription is ludicrous in these cases. How can a mousetrap *want* to trap a mouse? How can it *believe* a mouse is chewing on the bait when it *feels* the cheese move? Ascription is appropriate, it seems, only when the subject is a person.

Again the question is begged, are we more than a complex mousetrap? is there a fundamental difference in complexity that makes people what they are? or merely a quantitative one? One commonly presented difference is intelligence. Alan Turing believed that a valid test for intelligence was what he called the "Imitation Game" (aka Turing Test). If an AI were able to answer questions in a way that was indistinguishable from a person, then the AI can be deemed "intelligent" (Haugeland, p6). He and many others believed that intelligence was merely a matter of sufficient complexity. Strong AI, or "Computationalism", says that we will someday be able to construct a system able to pass the Turing Test (Bringsjord, p4). Once we are able to construct such a system it goes without saying that we will understand the rules that are followed by this system. If these rules turn out to be simple and explain all available evidence, some might be inclined consider our understanding our how the mind works as complete.

Is there any hope of understanding our mind so well? Are we perhaps fated to fail by our own patterns of thinking? Can we perhaps be trapped by our own questions? To form a question we must first put it into words. Benjamin Lee Whorf said that "We dissect nature along lines laid down by our native languages". Brian Arthur, a researcher

of Complexity Theory, says that all our intellectual pursuits are plagued by assumptions, and when we drop certain assumptions, the very questions we ask will change. We will no longer attempt to optimize an interpretation of a system in terms of one agent versus another, or even look for equilibrium, but will instead understand that everything in this world is ever-changing, nonlinear, and a participant in it (Waldrop, p333-5). If one can never step in the same river twice (Heraclitus) how can one study the same mind twice?

Complexity is the study of systems comprising agents or parameters so numerous that the rigorous modeling of such systems is mathematically or computationally infeasible using present methods (Waldrop). Is not the brain such a system? If each neuron is an *agent* in this system, we are hopelessly ill-equipped to rigorously model the action of each neuron in a person's brain. We are even less equipped to study the *mind*, if one considers that we do not even know what the *agents* of said system would entail.

Let's say for the sake of argument that we are one day able to determine the *atom* of thought - let's call it a logical *node*. Each node is an agent in the overall system of the subject's mind. In a best case scenario we will be able to describe a persons every observable behavior with a set of rules in terms of these nodes. For instance, Chomsky describes the Case-Filter as being a logical computation performed in the mind with the simple rule that every phonetically realized noun phrase (NP) must be assigned case, albeit abstract in some sentences (Chomsky, p74, 43). Stated in terms of a node ***C*** representing our ability to assign case to a noun phrase, where ***N*** is the phonetically realized noun phrase:

$$\textit{state of case assignment} = C(N)$$

In an example phrase such as "for there to be ...", ***there*** receives case from ***for*** (Chomsky, p94). An instance of the function ***C (N)*** could be represented (simplified for brevity):

$$C(\textit{there}) = x \quad \text{if} \quad \textit{there} \ \text{comes after} \ \textit{for}$$

To be sure, a rigorous definition of every linguistic rule in this manner would take eons - some have attempted it. The bad news from Gödel is that even if every rule governing our thought can be determined and represented in an interlocking set of Node Functions, unprovable statements would still exist. At best, the set of all Node Functions would comprise a formal system, and all formal systems have been proven to be incomplete. At worst, our set of functions would be *inconsistent*. If so, then the set has a chance of being complete, but an inconsistent system is arguably useless in the pursuit of science.

I would like to propose that a Node Function such as the one above can be termed an *understanding*. Let the writing of such a function be a rigorous scientific definition of "*understanding a mind process"*. When an observer is able to describe the mind process of a subject in such a way, the observer has *understood* the process within said subject. If we assume (however unlikely) that we could someday model every computation of the mind as Node Functions, it would also be possible to represent the Node Function for understanding a given Node Function, and the understanding of the understanding of a process, ... ad infinitum. How useful is such research? If our road will be blocked at some point by inherent incompleteness at best, or by inconsistency at worst, is the quest for an infinite set of mental rules really worth it? Even if they can be defined intensionally as a finite set, are we perhaps already "trapped" by our own assumptions?

An assumption made by Wittgenstein is that we each follow rules ourselves. He says that we can determine whether or not others think like us by interpolating their rules. These rules can be discovered by examining their actions. What if no one actually follows rules? If our minds are better described by the mathematics of Complexity, perhaps there is another explanation for language perception. Assuming that a speaker and a listener both have minds of similar levels of complexity, upon hearing a segment of speech, the listener can take perceptual auditory input and use his own complex system as an *emulator* to parse the speech. Perhaps rules would still have a place in this paradigm, but they would be relegated to the status of *approximations*. Chomsky certainly supports the idea of *emulation* by saying that we don't need to follow just one set of rules. We can understand a speaker even when the speaker applies a rule we personally do not, as long

as we have internalized that rule at some point (Chomsky, p242). In this case, we implicitly emulate the rule set of another person.

An assumption made by Chomsky is that we follow the simplest rules. He says that when multiple theories explain the observed data, the best theory is the one that incorporates the fewest details - therefore the simplest one (Chomsky, p255). If neural nets can be taken to be an appropriate analogy to some neurological structures, it would be important to note here that neural nets only arrive at the simplest solution under the best conditions. If the designer is sufficiently competent, a neural net will arrive at a configuration representing the lowest possible complexity. Otherwise, it will stop adjusting itself when it reaches a local minimum. The problem of local minima is inherent in any multidimensional math problem. When there is no designer, as Chomsky believes (Chomsky, p238), there is no scientific reason to assume that a complex system can arrive at the simplest solution. All things being equal, a local minimum is much more likely.

The theories laid out by Chomsky and Wittgenstein have both advanced scientific pursuits greatly. The value of their accomplishments are not being debated here. What is being presented are questions that need answers. Perhaps two new fields of study should be commenced: The study of Logical Node Functions within the mind could be termed, *Cognitive Mechanics*, and the study of macro-mental processes could be termed *Cognitive Complexity*. Could a mind ever understand itself? "Not completely", says Gödel. Descartes believed that there are some aspects of thought that may forever surpass human understanding, and it may not be because of limited human intelligence (Chomsky, p223). It may be that some aspects are beyond any intelligence the physical world could support.

SOURCES

Bringsjord, Selmer (September 24, 2001) "Strong AI/Computationalism Defined" RPI Course Material

Chomsky, Noam (printed 1986) "Knowledge of Language", Praeger Publishers

Haugeland, John (printed 1990) "Artificial Intelligence, The Very Idea", The MIT Press

Hofstadter, Douglas (printed 1999) "Gödel, Escher, Bach: an Eternal Golden Braid", Basic Books

Gödel, Kurt (printed 1992) "On Formally Undecidable Propositions of Principia Mathematica and Related Systems", Dover Publications Inc.

Lindblad-Toh, Kerstin, "Three's Company" Nature, vol 428, 1 April 2004

Pinker, Steven (printed 1997) "How the Mind Works", W W Norton & Company

Waldrop, M. Mitchell (printed 1992) "Complexity" Simon & Schuster, New York

Wittgenstein, Ludwig (printed 1960) "The Blue and Brown Books", Harper & Row