Theoretical Constructs and Measurement of Performance and Intelligence in Intelligent Systems

Larry H. Reeker
National Institute of Standards and Technology
Gaithersburg, MD 20899
(Larry.Reeker@NIST.gov)

Abstract

This paper makes a distinction between measurement at surface and deeper levels. At the deep levels, the items measured are theoretical constructs or their attributes in scientific theories. The contention of the paper is that measurement at deeper levels gives predictions of behavior at the surface level of artifacts, rather than just comparison between the performance of artifacts, and that this predictive power is needed to develop artificial intelligence. Many theoretical constructs will overlap those in cognitive science and others will overlap ones used in different areas of computer science. Examples of other "sciences of the artificial" are given, along with several examples of where measurable constructs for intelligent systems are needed and proposals for some constructs.

Introduction

There are a number of apparent ways and certainly many more not so apparent ways to measure aspects of performance of an intelligent system. There are a variety of things to measure and metrics for doing so being proposed at this workshop, and it is important to discuss them. To develop a measure of machine intelligence that is supposed to correlate with the system's future performance capability on a larger class of tasks considered intelligent would be analogous to human IQ. That would require agreement on one or more definitions of machine intelligence and finding a set of performance tasks that can predict the abilities required by the definition(s), and still might not say much about the nature of machine intelligence or how to improve it.

One reason that metrics of performance (and perhaps, of intelligence) are needed is that they directly address the fact that it has been difficult to compare intelligent systems with one another, or to verify claims that are made for their behaviors. Another reason is that having measurements of qualities of any sort of entity provides a concrete, operational way to *define* the entity, grounding it in more than words alone. All of these aspects - *comparability*, *verifiability*, and *operational grounding* - were undoubtedly at least part of what Lord Kelvin meant about measurements providing a feeling that one understood a concept in science. (See the preamble to this workshop [Meystel *et al* 00]: "When you can measure what you are speaking about and express it in numbers, you know something about it.")

The measurements that form the primary topic of this paper are of a different type. They are ones that look ahead to the future, when the intelligent systems or artificial intelligence* field is more mature. The notion of mature field is defined here in terms of scientific theories that predict the performance of the systems on the basis of the underlying science. It is suggested that really valuable measurements require reliable predictions of this scientific sort, rather than just ways to compare the technological artifacts based on the science. To do this, it is necessary to develop theories containing measurable theoretical constructs, as will be discussed below.

The discussion of metrics for attributes of theoretical constructs herein does not conflict in any way with the idea of overall system measurements, comparisons, or benchmarks, which are useful for the purposes mentioned above. In fact, it is a philosophical problem to decide where theoretical constructs stop and empirical constructs begin. Measurements of artifacts will be referred to as **surface measurements**, those of a more theoretical nature as **deep measurements**, terms borrowed from Noam Chomsky's [65] terms for levels of syntactic description. The question of "how deep" can be left open at this time. This paper advocates looking for measurable theoretical constructs at the deeper level that will predict surface behaviors at the level of the system or subsystem, or of an entire artifact.

* The latter term will be used herein because the shortened form, "AI" is more common than "IS".

The remainder of the paper explains the form that we will expect for AI theories in the future if they are to qualify as scientific theories and suggests theoretical constructs that may have measurable properties. It will discuss existing constructs that are developing as candidates for deep metrics and how they may relate to surface measurement. It will compare them to constructs in existing scientific theories at deep and surface levels. It will suggest that they will naturally relate to constructs from the artificial and natural sciences, specifically from cognitive science and computer science.

Computation Centered and Cognition Centered Approaches to AI

At all levels, from surface to deep, the constructs to be measured may depend on the approach taken to AI. There are two distinguishable approaches that have been taken over the years, which we will call "computation centered" and "cognition centered"*. The computation centered approach focuses on how certain tasks can be accomplished by artificial systems, without any reference to how humans might do similar tasks. We do not usually think of numerical calculation as AI, but if we did, we would have to think of the way it is done as computation centered. There is no particular reason to make it cognition centered.

In the cognition-centered approach to AI, the tradition is to discover human ways of doing cognitive tasks and see how these might be done by intelligent systems. Sometimes the motivation for this approach has been to try to find plausible models for human cognitive processes (cognitive simulation), but for AI purposes, it has often been a matter of using human clues to try to accomplish the computation centered approach. Some researchers feel that developing the artifacts using cognitive ideas may lead to more robust AI systems (using "robust" in the sense that the system is not narrow or "brittle" in its intelligent capabilities). But it is a natural way to think about the developing AI capabilities, since not all areas related to intelligent activities have been explored and reduced to mathematical methods to the extent of numerical calculations, or even of mathematical logic, which might directly facilitate a computation centered approach.

Mathematical logic makes an interesting case for pointing out that most AI researchers in practice blend the computation centered and cognition centered approaches, since it is formalized, yet still can be approached in a cognition centered way. Computers actually implement mathematical logic, which is essential in control statements of programming languages. However, actually proving theorems in logic (beyond propositional logic, where truth-table methods can be used), is a creative intelligent activity. There, things become more complex, in different ways. The first complexity is that is a *creative* activity and we do not really understand even how people do it. Secondly, it is *informationally complex*: there are inherent undecidability problems in logics of sufficient richness for most interesting purposes.

In attempts to make it easier for humans to prove theorems, natural deduction methods were invented by Gentzen [34] and developed by a number of people, notably Fitch [52]. In a sense, natural deduction can be thought of as a computation-oriented version of theorem proving, taking away some of the mental work of creativity. But this does not change the inherent informational complexity problems, which provide inherent limits on computability.

Going beyond logic to general problem solving one finds some empirical studies of effective ways in which humans do it that antedate the computer. One of them, means-ends analysis, was codified in the General Problem Solver (GPS) program of Newell and Simon. [63] (See also Ernst and Newell 65]. For programs in the GPS era, it was in the spirit of that work to attempt measurement of the extent to which the program could mimic human behavior. This was done by also studying verbal protocols of people solving the problem. Any way of comparing those to the performance of the program was still pretty much a surface measurement. Such surface measures of cognitive performance, are also the heart of the Turing test [Turing 50], but do not tell us much about what is happening deeper in the system, as Joseph Weizenbaum showed with Eliza [66] (and emphasized in an ironic letter [74]). In more recent times, case-based methods have been advocated [Kolodner 88] as relating to way some people solve problems and they do look

* In the email exchange leading up to the Workshop, a third approach, "Mimetic Synthesis", whose prime concern is the "Turing test" one of representing a computer to a human user as if it were another human, was distinguished from the two mentioned by Robby Garner. It is a good distinction, though like the others, the boundaries are not always clear.

very promising. Some of the constructs from these problem-solving methods will be mentioned below.

Though computation centered and cognitive centered approaches blend well, the measurements that occur to the developers in the two approaches will naturally differ, and this is particularly true as one tries to go to a deeper level by using constructs that are based either on cognition or on computation. In other words, AI may have measurable constructs coming from at least two different sources, the computation side and the cognitive side. This fact has some interesting implications as one looks at the measurement of deeper constructs, which may have to be reconciled with both approaches to be meaningful.

The Structure of Scientific Theories

Today's views of scientific theory have changed from those held in the 19th Century, Lord Kelvin's time. The bare-bones version of a scientific theory today is that it consists of a model composed of abstract **theoretical constructs** and a **calculus** that manipulates these constructs in a way that can account for observations and accurately predict the value of experiments. The model is as central today as was the notion of measurement to Kelvin. The theoretical constructs have a relation with observed entities, properties and processes that may be quite abstract, not necessarily readily available to human senses, but following directly from calculations based on the theory. There are a number of principles applied to a model that give us increased confidence in the theory, but the one most relevant here is that we can measure the observed entities to confirm the predictions of the theories. So Kelvin's concern has been preserved, but augmented, in today's view of theories.

It is relevant to observe that the "calculus" mentioned above is used in the dictionary sense "a method of computation or calculation in a special notation (as of logic or symbolic logic)". That means that it may be numerical or non-numerical. In fact, as Herb Simon and Allen Newell [65] pointed out, there is no reason that the calculus cannot be expressed in the notation of a computer program, the better to speed its manipulation of the theoretical constructs.

For scientific theories in AI to be respectable, there will be certain requirements on them, and these affect whether they are accepted or not and whether the theories in which they occur are accepted. The late Henry Margenau had a pragmatic treatment of these requirements in his book *The Nature of Physical Reality* [Margenau 50]. A working Physicist as well as a philosopher, Margenau stressed that no amount of empirical evidence was scientifically convincing by itself, since it did not specify a unique model; and he also stressed the need for the binding of theoretical constructs to one another in a "fabric". This fabric was made up of theory and of mappings to empirical data. The theory was convincing to the degree that certain criteria were met - not a "black and white" situation, but one of degree. One of the criteria was the extent to which the models and constructs were extensible to larger and larger areas of scientific endeavor. As the fabric of the theory became larger and stronger, it became more difficult to rip it asunder.

Perhaps our emphasis on finding metrics can solidify the theoretical constructs of the field, as well as providing a means of measuring progress. The key to doing this is not to think of evaluation only as measurement of some benchmarks or physical parameters ("behaviors") that are manifested in the operation of the systems being evaluated. We need to be thinking in terms of the inner workings of the systems and how the parameters within them relate to the measured externally manifested behaviors.

One of Lord Kelvin's special interests was temperature. Temperature is of course something that we experience, something not wholly abstract. Certain physical properties are related to temperature, and the most easily observed is freezing and boiling of water. It took some scientific discovery to realize that each of these phenomena always take place at a particular (with a few reservations, like altitude and purity of the water), but still, those are concrete embodiments. Temperature has been a subjective attribute during most of the history of mankind, but the scientific notion of temperature is a theoretical construct, even though it has a close correspondence to subjective experience. The particular metrics chosen related to water boiling (in both Fahrenheit and Celsius), to Freezing (in Celsius), and to the "coldest" temperature that could be achieved with water, ice and salt (in Fahrenheit). Lord Kelvin also took the amazing step of developing a notion of temperature that is *really* abstract. His zero point of minus 273.15 degrees Celsius has never quite

been reached, and is far below what any person could experience. Yet it is very real as a scientific construct, one that is part of the fabric of physical science and ties various aspects of science together in that fabric.

Many other common terms in physical theory, like mass and gravity, are theoretical constructs, though they are related to human senses. Only in relatively recent physics history have mass and gravity been understood, and we owe that understanding to bits of inspiration on the part of Galileo and Newton. Having only half a century of AI history to look back on, we cannot really expect to have such a firm fabric of theoretical constructs stitched together. But some ideas are given below, after a comparison of Sciences that study natural and the artificial systems.

Sciences of the Artificial and their relation to Natural Sciences

Herbert Simon came to the conclusion that there was a place for what he called "Sciences of the Artificial" in his important book [69]. He did not *invent* the study of artifacts in a systematic manner, but he realized accurately and acutely that that artifacts could be subjects of "real sciences", with deep theories of the sort that exist in natural sciences. We will now consider some of the implications of this idea.

The boundaries between sciences of the artificial and the natural sciences are not clear-cut in practice because nature colors human artifacts, determining their possibility and their features. The "engineering sciences", the portions of engineering that has been formalized in the sense of that they can predict the behavior of artifacts, including aspects such as stability and strength can be considered sciences of the artificial. The reason that this is not remarked upon more often is that they have called upon physical sciences more and more over the centuries to aid the "ingenuity" that gives the profession its name.

Linguistics is a science of the artificial. Human language is the artifact that it studies. But of course, the properties of the artifact are shaped by the natural properties of human learning and cognition, human hearing and speech in many ways. In the domain of phonetics, for example David Stampe's "natural phonology" [Stampe 73, Donegan and Stampe 79] characterizes some of the interactions between language as an artifact and as a natural phenomenon. We do not understand even yet the extent of the interaction between linguistics and human cognition. Is there an LAD (language acquisition device) [Chomsky 75] innate in humans that is specific to language, or is the learning of language based on the same principles as such other acquired systems as visual perception? Nobody knows for sure; but whatever the case, the nature of the world and the nature of learning processes must affect language.

Computer Science is a science of the artificial. Certainly, this is true insofar as it studies computers, which are artifacts; but also to the extent that it studies algorithms, which are human creations, too. The main subject studied in much of Computer Science is not computers but information, and the "state", which is all the relevant information about a system at a given time, is therefore a fundamental theoretical construct. Information is a theoretical construct that is also fundamental in the natural sciences, but whose significance as a theoretical construct has only become apparent in this century, as its relationship to entropy and its role in quantum theory have been realized. So again, Computer Science has both artificial and natural parts.

Economics, another science of the artificial, studies a major artifact, the economy, and looking at this science of the artificial can provide some insight into the position of AI as a science of the artificial, and of the role of measurable theoretical constructs.

Predictive Measurement in a Science of the Artificial – An Example from Economics

Economics has struggled for longer than AI has existed to find theoretical constructs that have predictive power. Economics deals with large amounts of aggregated data, so its empirical data are statistical in nature, and its models are not as clear as physical models with respect to the interrelationships among theoretical constructs, nor are they as widely accepted. Yet they do allow some prediction of economic performance and are used in control processes for the purposes of economic stability, with a degree of success.

As this paper is being written, the U.S. Federal Reserve Bank is aggressively raising interest rates because the *employment rate* (inferred from job creation and unemployment data) is high and *economic growth* (a function of GDP change and other data) has been rapid. In

their models, this predicts increasing *inflation* (as measured by the *consumer and producer price indices* and other constructs). It has recently been conjectured that there should be a role in these models for *productivity*, the role of which is not yet fully understood. So economic theory, as it develops, must relate all of these constructs and others: average interest rates, supply and demand for money and goods, savings rate, etc.

Economic theories and their constructs are still complex and incomplete. Incorporated in complex computer models, their predictions are not totally trustworthy, but the predictions are testable. Economics provides an example from another science of the artificial that AI should follow in formulating and measuring constructs.

Surface Measures and Theoretical Constructs in AI – Some Examples

The sort of predictive ability that economists want, we would like to see in AI, too. If we have theoretical constructs at some deeper level, we can also use the theories of which they are a part to simulate or predict mathematically what happens if we increase or decrease parameters related to those constructs. It is a thesis of this paper that *there are theoretical constructs that can predict system performance measured in terms of surface measures*. At this point in the development of AI as science, it is hard to say just exactly what they would be, but some ideas can be drawn from today's AI and related subjects.

An Example Construct: Robustness

A surface measurement that could be very valuable across a variety of systems is some measure of *robustness* – the ability to exercise intelligent behavior over a large number of tasks and situations. From a computation-centered standpoint, if systems become robust, AI progress would be easier to see. From a cognition-centered standpoint, a system can never really be intelligent if it is not robust. (One way to think of a measure of intelligence in a single system would be as a measure of performance, robustness and autonomy.) The *surface* way to determine the robustness of a system would be to try it on a number of tasks and see how broad its methods are. But what *makes* intelligent systems robust? Learning ability, experience, and the ability to transfer that experience to new situations are all things that come to mind. A rough sketch of how measuring theoretical constructs in those areas might give us a predictive figure for developing robust systems is given below.

Robustness: Learning?

If learning can make systems more robust, it should be interesting to measure the strength of the system's learning component. How easily does it adapt the system to a new situation? *Unsupervised learning* has wide applicability, but it can basically only determine clusters of similar items. *Supervised learning* must be presented with exemplars to learn relations, which seems not to be enough for a machine to extend its own capabilities. *Reinforcement learning* (RL) is a blend of both cognitive and computational centered AI. It started out as a model of classical conditioning, but turned out to be applied dynamic programming. There are a number of different techniques within RL, all of which have many possible applications. Neural nets or other approaches may be used. The theoretical constructs include the *state space* chosen, the *reinforcement function*, and the *policy*. The field is becoming quite sophisticated, and there are known facts about the relation of these to outcomes in particular cases [Mahadevan and Kaelbling 96]. Suppose that a reinforcement learning system constitutes a part of the intelligence of an intelligent system. There should be some way of predicting how that system would do upon encountering problems of a certain nature. By knowing how it chooses the concepts in its system and how they react on problems of that type, one can provide a partial evaluation of how effective the learning system would be. By obtaining such figures for all such subsystems, one could relate them to the performance of the full intelligent system. There is much work to be done in that direction.

Under certain circumstances, one can imagine learning extending robustness; but having to learn each new variations of a problem, even by reinforcement, is unlikely to lead to robustness quickly. It is expected that reinforced behaviors learned in one situation might be identical to those needed in another system, so this may lead to more rapid or better learning in the second situation. One approach to this is to condition behaviors that are not built into the system initially, as explored by Touretzky and Saksida [97]. But, still, one would like to have more general ways of reusing "big pieces" of learned knowledge.

Robustness: Transfer of Learning?

Transfer of learning is a phenomenon that we may be able abstract to theoretical constructs that can help to predict robustness. It is still not a deep measure, so it will then be important to predict transfer of learning from deeper constructs which will be mentioned below. At present, it is a research challenge to build transfer of learning into systems. But it is possible to see how one could test for it.

As far as measurement, here is roughly how transfer of learning might be measured:

1. Machine performance is measured on Task 1. The score is P(t1, T1)) = performance at time t1 on Task 1. P is some suitably broad performance measure.
2. Performance is measured on Task 2 without learning (this being an artifact where we can control learning) to obtain P(t1, T2) (keeping the time variable the same because the same machine abilities are assumed without learning even if the measurements are not simultaneous).
3. Note that if the measure is to have a meaning, previous training that might affect T1 or T2 must be controlled for, which could be difficult.
4. The machine is now allowed to perform task T1 in which it learns, achieving better performance at some time t2, i.e. P (t2, T1) > P (t1, T1).
5. It is then tested on T2, and the question is whether P (t2, T2) > P (t1, T2) without having done additional learning on Task 2.

If indeed P (t2, T2) > P(t1, T2) in some quantifiable way, the system has achieved (at least locally) one of the goals of AI, the transfer of learning from T1 to T2. The amount of transfer can be measured by the amount of improvement on task2 as a function of the amount of training on task T1. Let us assume that we can describe this by some transfer effectiveness function, E for the system being tested. Let us say E(T1, T2, t) gives "the effectiveness of training on T1 for time t in terms of transfer toT2". We could describe this by a graph of performance on T2 as a function of time being spent on T1.

Developing such a measure of transfer of learning and getting it accepted is not simple. To be useful, we would need a way of comparing T1 and T2, to be sure that the second task is not just a subtask to the first. Difficult or not, defined measurements such as these are steps toward understands the construct "transfer of learning" and achieving it in artifacts. The measurable transfer construct would, in turn, help to provide a measurement of robustness, since learning transfer can make a system more robust. It is a step toward measurement of intelligence, at least by some definitions of intelligence, and, intuitively, at least, would have some predictive power.

How might we go about defining the similarity of T1 and T2, as suggested above? We would have to decide what we mean by similarity of task. An interesting essay in this area is "Ontology of Tasks and Methods" [Chandrasekaran, Josephson and Benjamins [98]].

Various candidates for potentially measurable constructs that could be used to produce transfer but also to relate transfer to other phenomena are mentioned in a book edited by Thrun and Pratt [98], who have both had a research interest in learning-transfer processes. From the computation side comes the possibility of changing *inductive bias*. From the cognition-centered side, there is *generalization* from things already learned; but *overgeneralization* can be a major problem in learning, so it needs to be constrained. (Some simple constraints on overgeneralization in language learning are discussed in [Reeker 76].)

Robustness: Case-Based Reasoning?

Case-based reasoning is an intuitively appealing technique that was mentioned earlier in this paper. The idea is that one learns an expanding set of cases and stores the essentials of them away according to their conventional features. They are then retrieved when a similar case arises and mapped into the current case. Potential theoretical constructs include *indexing* and *retrieval* methods for the cases, case *evaluation* and case *adaptation* to the new situation. The cases could also be abstracted and generalized to various degrees, to a *model.*

Case-based reasoning is important for cognition centered AI. It is intuitively the way many people often figure out how to do things, and is thus embodied in the teaching methods of many professional fields – law, business, medicine, etc. It provides a launching pad for creativity as well, as mappings take place from one case to an entirely new one. Perhaps the new case is not really concrete, but a vague new

idea. Then the mapping of an old case to it may result in a creative act – what we usually call *analogy*. Analogy, *metaphor* in language, is a rich source – absolutely ubiquitous – of new meanings for words, and thus of new ways to describe concepts, objects, actions. Perhaps one key to robustness is the ability to use analogy. Four interesting papers by researcher in the area can be found in an issue of *American Psychologist* [Gentner *et al* 97].

Existing Surface and Subsurface Performance Measures

Researchers in text-based information retrieval (IR) have traditionally considered themselves not to be a part of the AI field, and some have even considered that artificial intelligence was a rival technology to theirs; but there is an overlap of interest. It is worth noting that IR has had a useful surface measure of system performance that has guided research and allowed comparison of technologies. The measure consists of two numbers, *recall* and *precision* [Salton 71]. Recall measures the completeness of the retrieval process (the percentage of the relevant documents retrieved). Precision measures the purity of the retrieval (the percentage of retrieved documents judged relevant by the people making the queries). If both numbers were 100%, all relevant documents in a collection would be retrieved and none of the irrelevant ones. Generally, techniques that increase one of the measures decrease the other. Real progress in the general case is achieved if one can be increased without decreasing the other.

For the IR community, better recall and precision numbers have both shown the progress of the field. They also show that it is still falling short, keeping up the challenge, especially as the need to use it for very large information corpora rises. In addition, they provide a standard within the community for judging various alternative schemes. Given a particular text corpus, one can consider various weighting schemes, use of a thesaurus, use of grammatical parsing that seeks to label the corpus as to parts of speech, etc., to improve the retrieval process. The interesting thing is to relate these methods and the characteristics of the corpus to precision and recall, but so far that has not been sharp enough to quantify generally.

Related to information retrieval is automated natural language information extraction, which tries to find specified types of information in bodies of text (often to create formatted databases where extracted information can be retrieved or mined more readily). A related but different (cost-based) measure was defined several years ago for a successful information extraction project [Reeker, Zamora and Blower 83]. One measure was *robustness* (over the texts, not different tasks as in the broader intelligent systems usage discussed earlier). This was defined as the percentage of documents out of a large collection that could be handled automatically. The idea was that some documents would be eliminated through automated pre-screening (because those documents were not described by the discourse model the system used) and relegated to human processing. Another measure was *accuracy* (the percentage of documents not eliminated that were then correctly processed in their entirety, by the system). Yet another was *error rate* (the percentage of information items that were erroneous – including omitted - in incorrectly handled documents). From this more detailed breakdown, estimates of the basic cost of processing the documents, based on human and machine processing costs and costs assigned to errors and omissions, was derived. The measure could be used to drive improvements in information extraction systems or decide whether to use them, compared to human extraction (which also has errors) or to improve the discourse model to handle a larger portion.

For information extraction projects, it was further suggested that the cost of erroneous inputs might drive a built-in "safety factor" that could be varied for a given application [Reeker 85]. This safety factor was based on linguistic measures of the text (in addition to the discourse model) that could cause problems for the system being studied. The adjustable safety factor could be built into the prescreening mentioned above. In other words, the system would process autonomously to a greater or lesser degree and could invite human interaction in applications where the cost of errors was especially high. It was suggested that the system would place "warning flags" to help it make a decision on screening out the document, and these could also aid the human involved. Although this was a tentative piece of work, the idea of tying a surface measure (robustness) into the underlying properties of the system is exactly like tying measurable surface properties into underlying theoretical constructs. The theoretical constructs mentioned in this case were structural or semantic ones from linguistics.

From the area of software engineering comes another tradeoff measure that is worth mention. The author did some work on ways of providing metrics - surface metrics, initially - for program readability (or understandability) [Reeker, 79]. Briefly, studies of program understanding had identified both go-to statements and large numbers of identifiers (including program labels) as problems. At the same time, the more localized loop statements could result in deep embeddings that were also difficult to understand for software repair or modification. The vague concept of readability could be replaced by a measure of go-to statements and maybe also one of the number of different identifiers. This particular study suggested *depth of embedding* as a problem and also suggested a tradeoff between depth of embedding a metric called *identifier load*. Identifier load was a function of the number of identifiers and the span of program statements over which they were used. Identifier load tended to increase as depth of embedding was reduced by the obvious methods.

There were a number of similar software metrics studies in the 1970s, and they continue. This approach, however, was part of an attempt to look at natural language for constructs that might be of relevance in programming languages and programming practice [Reeker 80]. The *depth* measure was based on an idea of Victor Yngve [60], which came out of his work in linguistics - an idea that retains a germ of intuitive truth. Yngve had in turn related his natural language measure of embedding depth to measures of short-term memory from cognitive psychology. Whether these relationships turn out to be true or lead to related ideas that are true or not, they illustrate how theoretical constructs can stitch AI, computer science, and other artificial and natural sciences together. They also illustrate the quest for metrics that can firm up the foundations of the sciences.

More Constructs To Be Explored

There are many more existing theoretical constructs that have arisen within AI or been imported from computer science or cognitive science that beg to be better defined, quantified, and related to other constructs, both deep and surface.

Means-ends analysis and *case based reasoning* have both been mentioned as forms of problem solving. How do these cognitive characterizations of problem solving relate to one another? At a deeper level is the construct of *short term memory* mentioned in the previous section in relationship to Yngve's *depth*. How does short-term or working memory relate to long term memory and how are the two used in problem solving? The details are not known. The size of a short-term memory may not be as relevant in a machine, where memory is cheap and fast. But we cannot be sure that it is not relevant to various aspects of machine performance because it is reflected at least in the human artifacts that the machine may encounter. For instance, in resolving anaphora in natural language the problem may be complicated if possible referents are retrieved from arbitrarily long distances.

A similar problem arises from long-term memory if everything ever learned about a concept is retrieved each time the concept is searched for. This can lower retrieval precision (to use the term discussed earlier for machine retrieval) and cause processing difficulties on a given problem. It may be that Simon's notion of *bounded rationality* is a virtue in employing intelligence. Are we losing an important parameter in intelligence if we try always to optimize rationality? For AI system, *anytime algorithms* and similar constructs for approximate, uncertain, and resource bounded reasoning have been developed in recent years, and hold a good deal of promise [Zilberstein 96].

An interesting theoretical construct arising out of AI knowledge representation and the attempts to use it in expert systems and agents and for other purposes is that of an *ontology*. "Ontology" is an old word in philosophy designating an area of study. In AI it has come to designate a type of artifact in an intelligent system: The way that that system characterizes knowledge. In humans, ontologies are shared to a large degree, but certainly differ from every person to every other, despite the fact that we can understand each other. Are some ontologies indicative of more intelligence than others in ways that we can measure? One suggested criterion for high intelligence is the ability to understand and use very fine distinctions (or to actually create new ones, as described in Godel's memorandum cited by Chandrasekaran and Reeker [74]). Is an ontology's size important, or its organization, or both? Can one quantify a system's ability to add new distinctions?

A related issue is *vocabulary*. Many people think that an extensive vocabulary, *used appropriately*, is a sign of intelligence, or at least

Characterizing Three Related Endeavors Involving Computers and Intelligence and Their Purposes[2]			
Name of Endeavor	**Mimetic Synthesis**	**Cognitive Modeling**	**Artificial Intelligence**
Principal Goal	Produce behavior that appears to be evidence of cognition or intelligent phenomena	Produce models of cognitive processes, including learning, planning, resoning, perception, linguistic behavior, etc.	Find ways of doing with computers things that we deem intelligent when they are done by humans.
Use	Produce illusion of intelligent behavior for interface purposes, entertainment, etc.	Develop psychological theories of cognition.[1]	Tools to augment intelligence and systems that exhibit increasingly intelligent behavior autonomously.
Category of Endeavor	Computer Technology	Psychological Science (Branch of Natural Science)	Computer Science (Branch of Science of Artificial)
Approach	Use simulations, stored answers, AI or cognitive models, or other techniques that are convincing to human users.	Use evidence from psychological experiments; make working models; test against human behavior.	Use techniques from mathematical and engineering disciplines, cognitive models, and previous experience. Test through programs.
Examples	Eliza (J. Weizenbaum), Albert (Garner and Henderson), Talking Coke Machines (?),…	LT, GPS (Simon and Newell), HAM (J. Anderson), SOAR (A. Newell),…	Deep Blue (IBM), SATPLAN (Kautz and Selman), Dendral (Buchanan, Feigenbaum..), TD-Gammon (Tesauro),…

1. By this we mean the traditional cognitive psychology level, not brain function. The latter is a biological approach. In the nature of science, of course, one expects theories of such close areas to be consistent and to inform one another, and to merge in the longer term.
2. Clearly, each of these endeavors is different, though each can make use of knowledge from the others and some devices could solve all goals. The confusion between them, however, has resulted in misunderstandings for decades.

scholastic aptitude. In computer programs that do human language processing, the vocabulary consists of a *lexicon* that generally also has structural (*syntactic*) information for parsing or generating utterances containing the lexical item and *meaning representations* for the lexical item. The lexicon can be much larger than any human's vocabulary; but for the vocabulary to be used appropriately for language production or understanding, it still falls far short of the human vocabulary. For that to be improved better techniques of *semantic mapping* are required, including links to ontologies and methods of inferring the ontological connections and of idiosyncratic aspects of speakers with which a conversation is taking place. Is the vocabulary an indication of the size of the ontology and the distinctions it makes, or vice-versa? Nobody knows; but better theories of how they link up are needed for both understanding and fully effective use of human language by intelligent systems.

Another cognitive concept that is still a mystery is *creativity*, certainly a part of intelligence, or at least of high intelligence. Does the ability to add entirely new concepts, not taught, constitute creativity? How does one harness serendipity to develop creativity? Is creativity linked with *sensory cognition*, the cognitive phenomena related to senses, such as vision, including perception, visual reasoning, etc. There is a need for deep theoretical constructs underlying notions like creativity, and for measures of these constructs and their attributes [Simon 95, Buchanan 00].

Turning to computational constructs, we notice that much of the AI described above takes place through various forms of *search*. Already there exists a pretty good catalogue of variations on search and how to manage it, in which a good deal of theory is latent. Some of the search is of a *state space*, involving the ubiquitous state concept basic to theoretical computer science. Search is also coupled with *pattern matching*, which underlies many of the methods mentioned earlier in this paper.

The potential constructs mentioned here are just a sample of the ones already available in Artificial Intelligence, and to them should be added others found in some of the major works of Newell and Simon on Problem Solving and Cognition [Newell and Simon [65], Newell [87]].

Summary and Author's Note

The development of a true science of artificial intelligence is something that has concerned the author for a long time. It has been encouraging to see the development within the field of interesting and non-obvious theoretical constructs. This paper has suggested that theoretical constructs with attributes that we can measure are especially valuable and it has suggested a number of such candidates. The paper suggests that we enlist Lord Kelvin's emphasis on measurement in choosing such constructs. These same measurable theoretical constructs will in many cases relate (at least at deeper levels) to those of cognitive science, computer science, and other sciences. They will help predict measures at the surface that can be used to provide metrics for the performance (and through that, the intelligence) of intelligent artifacts. We should have in mind the quest for such measurable constructs as we move forward in creating intelligent artifacts.

References

Buchanan, B. G. [00], Creativity at the Meta-Level, Presidential Address, American Association for Artificial Intelligence, August 2000. [Forthcoming in *AI Magazine*.]

Chandresekaran, B., J. R. Josephson and V. R. Benjamins [98] Ontology of Tasks and Methods, 1998 Banff Knowledge Acquisition Workshop. [Revised Version appears as two papers "What are ontologies and why do we need them?," *IEEE Intelligent Systems*, Jan/Feb 1999, 14(1); pp. 20-26; "Ontology of Task and Methods," *IEEE Intelligent Systems*, May/June, 1999.]

Chandrasekaran, B. and L. H. Reeker [74]. "Artificial Intelligence – A Case for Agnosticism," *IEEE Trans. Systems, Man and Cybernetics*, January 1974, Vol. SMC-4, pp. 88-94.

Chomsky, Noam [65]. *Aspects of the Theory of Syntax*. MIT Press, Cambridge MA.

Chomsky, N. [75] *Reflections on Language*. Random House, New York.

Donegan, P. J. & D. Stampe [79]. The study of Natural Phonology. In Dinnsen, Daniel A. (ed.). *Current Approaches to Phonological Theory*. Indiana University Press, Bloomington, 126-173.

Ernst, G. & Newell, A. [69]. *GPS: A Case Study in Generality and Problem Solving*. Academic Press, New York.

Fitch, F. B. [52]. *Symbolic Logic*. Roland Press, New York.

Gentner, D., K. Holyoak *et al* [97]. Reasoning and learning by analogy. (A section containing this introduction and other papers by these authors and A. B. Markman, P. Thagard, and J. Kolodner, *American Psychologist*, 52(1), 32-66.)

Gentzen, G. [34] Investigations into logical deduction. *The Collected Papers of Gerhard Gentzen*, M. E. Szabo, ed. North--Holland, Amsterdam, 1969. [Published in German in 1934.]

Kolodner, J.L. [88] Extending Problem Solving Capabilities Through Case-Based Inference, In, *Proceedings of the DARPA Case-Based Reasoning Workshop*, Kolodner, J.L. (Ed.), Morgan Kaufmann, Menlo Park, CA.

Leake, D. B. Ed. [96] *Case-Based Reasoning: Experiences, Lessons, And Future Directions* Indiana University, Editor 1996, AAAI Press/MIT Press, Cambridge, MA.

Margenau, H. [50]. *The Nature of Physical Reality*, McGraw-Hill, New York.

Mahadevan, S. and L. P. Kaelbling [96] The National Science Foundation Workshop on Reinforcement Learning. *AI Magazine* 17(4): 89-93.

Meystel, A. *et al* [00] Measuring Performance of Systems with Autonomy: Metrics for Intelligence of Constructed Systems. *In this volume*.

Newell, A. [87] *Unified Theories of Cognition*. Harvard University Press. Cambridge, Massachusetts, 1990. [Materials from William James Lectures delivered at Harvard in 1987.]

Newell, A., and H.A. Simon [63] GPS, a program that simulates human thought, *Computers and Thought*, E.A. Feigenbaum and J. Feldman (Eds.), McGraw-Hill, New York.

Newell, A., & Simon, H.A. [65]. Programs as theories of higher mental processes. R.W. Stacy and B. Waxman (Eds.), *Computers in biomedical research* (Vol. II, Chap. 6). Academic Press, New York.

Newell A. & Simon H.A. [72] *Human Problem Solving*, Prentice-Hall, Englewood Cliffs, NJ.

Reeker, L. The computational study of language acquisition, *Advances in Computers*, 15 (M. Yovits, ed.), Academic Press, 181-237, 1976.

Reeker, L. [79] Natural Language Devices for Programming Language Readability; Embedding and Identifier Load, *Proceedings, Australian Computer Science Conference*, Hobart Tasmania.

Reeker, L., E. Zamora and P. Blower [83] Specialized Information Extraction: Automatic Chemical Reaction Coding from English Descriptions, *Proceedings of the Symposium on Applied Natural Language Processing*, Santa Monics, CA, Association for Computational Linguistics, 1983.

Reeker, L. H. [80] Natural Language Programming and Natural Programming Languages, *Australian Computer Journal* 12(3): 89-93.

Reeker, L. H. [85] Specialized information extraction from natural language texts: The "Safety Factor", *Proceedings of the 1985 Conference on Intelligent Systems and Machines*, 318-323, Oakland University, Michigan, 1985.

Salton, G. [71] *The SMART Retrieval System*, Prentice-Hall, Englewood Cliffs, NJ.

Simon, H. A. [69] *The Sciences of the Artificial*. Third Edition. Cambridge, MA, MIT Press, 1996. [Original version published 1969].

Simon, H. A. [95] Explaining the Ineffable: AI on the Topics of Intuition, Insight and Inspiration. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, IJCAI 95, Montréal, Morgan Kaufmann, Menlo Park, CA, Volume 1, 939-949.

Stampe, D. [73] *A Dissertation on Natural Phonology*. New York: Garland Publishing, 1979. [Original University of Chicago dissertation submitted in 1973.]

Thrun, S. and L. Pratt (eds.) [98]. *Learning To Learn*. Kluwer Academic Publishers.

D.S. Touretzky and L.M. Saksida [97] Operant conditioning in Skinnerbots. *Adaptive Behavior* 5(3/4):219-247.

Turing, A. M. [50]. "Computing Machinery and Intelligence." *Mind* LIX (236 ; Oct. 1950): 433-460 reprint in [*Collected Works of A. M. Turing* vol. 3: Mechanical Intelligence, D. C. Ince ed., Elsevier Science Publishers, Amsterdam, 1992: 133-160].

Weizenbaum J [66]. ELIZA - a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9, 36-45.

Weizenbaum, J. [74]. Automating psychotherapy. *Communications of the ACM*, 17(7):425, July 1974.

Yngve, V [60] The depth hypothesis, *Proceedings, Symposia in Applied Mathematics*, Vol. 12: Providence, RI, American Math. Society, 1961. [Based on publication under another title, 1960.]

Zilberstein, S. [96] Using Anytime Algorithms in Intelligent Systems, *AI Magazine*, 17(3):73-83, 1996.