A complete theory of tests for a theory of mind must consider hierarchical complexity and stage

Michael Lamport Commons and Myra Sturgeon White

Department of Psychiatry, Harvard Medical School, Massachusetts Mental Health Center, Boston, MA 02115-6113. **Commons@tiac.net**
mswhite@fas.harvard.edu http://www.tiac.net/~commons/

Abstract: We distinguish traditional cognition theories from hierarchically complex stacked neural networks that meet many of Newell's criteria. The latter are flexible and can learn anything that a person can learn, by using their mistakes and successes the same way humans do. Shortcomings are due largely to limitations of current technology.

Anderson & Lebiere (A&L) raise important issues concerning criteria for evaluating the cognitive theories on which computational systems designed to simulate human intellectual abilities are based. Typically, cognitive theories are indirectly evaluated based on a theory's capacity to be translated into a computational system that produces correct answers or workable rules. The Newell 12-Criteria Test (1992; Newell & Simon 1963/1995) that A&L propose to measure theories with, makes an important move towards measuring a theory's capacity to exhibit underlying behaviors supporting the expression of human cognitive processes.

We suggest a further dimension. Most cognitive theories are, like Athena, born fully formed, modeling the highest stages of development. However, human cognition is a product of developmental process. Humans learn to act by building one stage's actions on actions from previous stages, creating the capacity to perform ever more complex behaviors. Thus, to fully explain or model human intellectual capacity, hierarchical complexity must be factored into a theory. The *Model of Hierarchical Complexity* (MHC) (Commons et al. 1998) delineates these developmental changes (see Dawson 2002 for validity and reliability).

MHC identifies both sequences of development and reasons why development occurs from processes producing stage transition. It may be used to define complex human thought processes and computer systems simulating those processes. With this model, performed tasks are classified in terms of their order of hierarchical complexity using the following three main axioms (Commons et. al 1998). Actions at a higher order of hierarchical complexity

1. Are defined in terms of lower order actions;
2. Organize and transform lower stage actions;
3. Solve more complex problems through the nonarbitrary organization of actions.

The order of the hierarchical complexity of a task is determined by the number of its concatenation operations. An order-three task action has three concatenation operations and operates on output from order-two actions, which by definition has two concatenation operations and operates on an order-one task action. Increases in the hierarchical complexity of actions result from a dialectical process of stage transition (Commons & Richards 2002).

To stimulate human intellectual capacities in computer systems, we design stacked neural networks that recapitulate the developmental process. This approach is necessary because currently we lack the knowledge to build into systems the myriad key behaviors formed during the developmental processes. Moreover, we lack the technology to identify the intricate web of neural connections that are created during the developmental process.

These stacked neural networks go through a series of stages analogous to those that occur during human intellectual development. Stages of development function as both theory and process in these systems. Actions (i.e., operations performed by networks resulting in a changed state of the system) are combined to perform tasks with more complex actions, permitting the performance of more complex tasks and thereby scaling up the power. The number of neural networks in a stack is the highest order of hierarchical complexity of task-required actions identified by the model. An example of a six-stage stacked neural network based on the model of hierarchical complexity (Table 1) follows.

Table 1 (Commons & White). *Stacked neural network (example of Model of Hierarchical Complexity)*

Order of Hierarchical Complexity	What It Uses	What It Does
0. Calculatory	From Humans	Calculates and executes human written programs.
1. Sensory and motor	Caller's utterances	A front-end speech recognition system translates customers' utterances into words. These "words" serve as simple stimuli to be detected.
2. Circular sensory motor	Words from speech recognition system	Forms open-ended classes consisting of groups of contiguous individual words.
3. Sensory-motor	Grouped contiguous speech segments	Labels and maps words to concepts. Networks are initially taught concepts that are central to the company environment: Products and departments such as customer service, billing, and repair.
4. Nominal	Concept domains	Identifies and labels relationships between concept domains. Possible interconnections are trained based on the company's functions, products, and services. Interconnections are adjusted based on system success.
5. Sentential	Joint concept domains	Forms simple sentences and understands relationships between two or more named concepts. Finds possible locations to send customers' calls. Constructs statement on whether they want to be transferred to that department. Customers' acceptances or rejection feeds back to lower levels.

Example. A system answers customer telephone calls, transferring them to the proper area within a large organization. Transfers are based on the customer's oral statements and responses to simple questions asked by the system. The system is capable of a three-year-old's language proficiency. A front-end recognition system translates customers' utterances (system inputs) into words that will serve as simple stimuli. It also measures time intervals between words.

Stacked neural networks based on the MHC meet many of Newell's criteria. They are flexible and can learn anything that a person can learn. They are adaptive because their responses are able to adjust when stimuli enter the stack at any level. They are dynamic in that they learn from their mistakes and successes. In the example, the system adjusts the weights throughout the stack of networks if a customer accepts or rejects the selected neural network location. Knowledge integration occurs throughout the networks in the stack. Moreover, networks based on the MHC learn in the same way as humans learn.

Some criteria are less easily met. Given current technology, neural networks cannot function in real time, are unable to transfer learning despite abilities to acquire a vast knowledge base, and cannot exhibit adult language skills. Whether we can build evolutions into systems – or even want to – is open to question. Finally, given our current limited understanding of the brain, we can only partially emulate brain function.

Criteria and evaluation of cognitive theories

Petros A. M. Gelepithis

Cognitive Science Laboratory, Kingston University, Kingston-upon-Thames, KT1 2EE, England. **Petros@kingston.ac.uk**

Abstract: I have three types of interrelated comments. First, on the choice of the proposed criteria, I argue against any *list* and for a *system* of criteria. Second, on grading, I suggest modifications with respect to consciousness and development. Finally, on the choice of "theories" for evaluation, I argue for Edelman's theory of neuronal group selection instead of connectionism (classical or not).

Introduction. Anderson & Lebiere's (A&L's) target article is a useful contribution on the necessity and grading of criteria for a cognitive theory and their application of the Newell Test to classical connectionism and ACT-R a worthwhile exercise. The following comments are partly a criticism on their proposed list of criteria, partly a response to their invitation for modifications of their proposed grading, and partly a critique of their choice of theories for evaluation.

On the choice of criteria for a Theory of Mind (ToM).[1] A&L state that "[t]wice, Newell (1980; 1990) offered slightly different sets of 13 criteria on the human mind" and a bit further down that their table "gives the first 12 criteria from [Newell's] 1980 list, which were basically restated in the 1990 list" (target article, sect. 1: Introduction, para. 1). Neither of these two statements is correct (as Table 1 confirms).

Furthermore, A&L's list is closer to Newell 1980 than to Newell 1990. No justification for this proximity is provided. Given that Newell's (1990) seminal book is incomparably more comprehensive than his 1980 paper, one wonders about the reasons for A&L's choice. Clearly, their claim of having *distilled* (emphasis added) Newell's two lists (cf. target article, Abstract) cannot be justified either. Although I agree that A&L's list is adequate to avoid "theoretical myopia" (Introduction, para. 2), it will create distortions in our quest for a ToM on account of being restricted to a fundamentally impoverished coverage of human phenomena (excluding, e.g., emotion, creativity, social cognition, and culture). It is worth noting that although Newell (1990, sect. 8.4) considered the extension of a unified theory of cognition (UTC) into the social band an important measure of its success, A&L chose to exclude from their list the one constraint with a social element that Newell had included (see item 9 in Table 2).

In contrast, evolution should not be a criterion! Humans are physical objects, but biology is fundamentally different from physics. Similarly, humans are biological systems, but psychology is fundamentally different from biology. The nature of human understanding (Gelepithis 1984; 1991; 1997) transcends the explanatory framework of modern Darwinism and, most importantly, of any future evolutionary theory. (For similar conclusions drawn upon different premises, see Mayr 1988; O'Hear 1997.)

Finally, a fourth list – very different from all previous three – has been offered by Gelepithis (1999). Of the four proposed lists, Table 2 juxtaposes the latest three. The reader can easily spot a number of obvious and significant differences among the three lists. For some of the less obvious, their corresponding serial numbers are in boldface. What all three have in common is that they do not provide necessary and sufficient conditions for a ToM. Still, the mind is a system (Bunge 1980; Hebb 1949; Sherrington 1906). We need, therefore, a *system* (not a list) of criteria characterising mind. A recent promising effort along this route is exemplified by Gelepithis (2002), which presents an *axiomatic system* delineating the class of intelligent systems as a foundation for the development of a ToM[2].

On some "objective measures." *Consciousness.* There are many volumes of readings (e.g., Hameroff et al. 1998; Revonsuo & Kampinnen 1994; Velmans 1996) at least as good as the one cited by A&L. Suggestions of measures on the basis of consciousness-related phenomena in one volume of readings should be avoided. Although universal agreement on what constitutes consciousness is nonexistent, Gelepithis (2001) has provided a list of "topics that, *presently,* constitute the major issues in the study of consciousness." I propose that list as a measure.

Table 1 (Gelepithis). *Extent of the overlap among the proposed sets of criteria by Newell and A&L*

Criteria	Comparisons with Respect to Newell's 1980 List		Comparison with Respect to Newell's 1990 List
	Newell 1990	**A&L 2003**	**A&L 2003**
New criteria	2	0	0
Significantly different criteria	3	2	5 or 6
Essentially equivalent criteria	3	3	3 or 2
Identical criteria	5	7	4