having system with a prewired repertoire of reflexes (unconditioned stimuli and responses – US, UR) that enable it to interact with its environment and accomplish simple automatic behaviors. The activation of any reflex, however, also provides cues for learning that are used by the adaptive control layer via representations of internal states. Adaptive control provides the mechanisms for the adaptive classification of sensory events (conditioned stimulus – CS) and the reshaping of responses (conditioned responses – CR) supporting simple tasks, and can be seen as a model of classical conditioning. The sensory and motor representations formed at the level of adaptive control provide the inputs to the contextual control layer that acquires, retains, and expresses sequential representations using systems for short- and long-term memory. The contextual layer describes goal-oriented learning as observed in operant conditioning. Central-processing steps at this level in the architecture are the following: (1) The representations of sensory cues (Sns) and associated motor states (Act) acquired by the adaptive layer are stored in short-term memory (STM) as a segment. (2) If a goal state is reached, that is, a target found or a collision suffered, the contents of STM are retained in long-term memory (LTM) as a sequence. Each segment of LTM consists of a sensori-motor representation (Sns, Act) a trigger unit (black) and a collector unit (white). (3) The reactive and adaptive control layers can still trigger actions and stand in a competitive relation to the contextual control system. (4) Each Sns state of the adaptive layer is matched against those stored in LTM. (5) The collector units of LTM can trigger actions dependent on the biased competition between LTM segments. By modulating dynamic thresholds of each LTM segment, different chaining rules can be implemented.

In panel B of Figure 1, the robot learns to use the color information in the environment, the patches on the floor and the walls, in order to acquire the shortest route between goal locations, that is, light sources (grey circles). The trajectory visualized is generated during a recall task where the light sources are switched off, after learning for about 30 min. The environment measures about 1.5 by 0.8 m; and the robot, about 55 by 30 mm.

A multilevel approach to modeling human cognition

Hongbin Wang,[1] Todd R. Johnson, and Jiajie Zhang

School of Health Information Sciences, University of Texas Health Science Center at Houston, Houston, TX 77030. **hongbin.wang@uth.tmc.edu todd.r.johnson@uth.tmc.edu jiajie.zhang@uth.tmc.edu http://www.shis.uth.tmc.edu**

Abstract: Although we agree with Newell and Anderson & Lebiere (A&L) that a unified theory of cognition is needed to advance cognitive science, we disagree on how to achieve it. A hybrid system can score high in the Newell Test but may not offer a veridical and coherent theory of cognition. A multilevel approach, involving theories at both psychological and brain levels, is suggested.

Newell certainly had a very good reason for being frustrated over the progress toward a scientific understanding of the human mind. The human mind is undoubtedly one of most complex entities in the world. It is systematically shaped by genetic and evolutionary forces; fundamentally constrained by physical and biochemical laws; influenced by cultural, social, and environmental factors; and manifests itself both psychologically and neurophysiologically. Given its inherent complexity and our limited knowledge in each of these aspects, it is conceivable that we may not be able to achieve a thorough understanding of the mind's work for a long time.

While we share Newell's frustration, we doubt that the Newell Test, as proposed in the target article, would offer us relief. On the one hand, the attainability of the test is theoretically questionable. It remains controversial, for example, whether self-awareness and consciousness are computationally implementable (e.g., Penrose 1989; 1996; 1997). This controversy helps to explain why both connectionism and ACT-R were graded "worse" on criterion 8 (self-awareness and consciousness) in the target article. On the other hand, even if we ignore the possible theoretical difficulties, we may still encounter practical problems in developing theories of mind that can pass the test, as we elaborate later.

After evaluating connectionism and ACT-R based on the Newell Test and suggesting that neither was satisfactory on all criteria, the authors Anderson & Lebiere (A&L) go on to recommend some remedies. One major remedy suggested is that we should somehow dissolve the distinctions and join the two approaches close together. Specifically, ACT-R needs to be "more compatible with connectionism," and connectionism needs to be concerned "with more complex tasks and symbolic processing" (sect. 6, para. 3). The authors note that building hybrid systems that can integrate the two approaches is particularly promising (ACT-R itself is already a form of hybrid system). By combining the advantages of different sub-approaches, the authors seem to suggest that hybrid systems would bring us one step closer to a Theory of Mind (ToM) that can score high in the Newell Test.

Unfortunately, there are at least three problems with this hybrid system approach. First, it should be noted that there are two (out of 12) criteria on which both connectionism and ACT-R score worse or worst. They are criterion 8 (self-awareness and consciousness) and criterion 11 (evolution). The simultaneous failure of both approaches on both criteria suggests that simply hybridizing the two approaches might not provide a solution.

Second, what if we develop a theory of self-awareness and an evolutionary ToM, and then hybridize these two theories with the hybrid system we constructed earlier? Does this give us a better ToM? Well, maybe. If doable, it will certainly boost the Newell Test score! But it also induces a paradox. Focusing on isolated and segmented subtheories of mind is what frustrated Newell and motivated the creation of the Newell criteria in the first place. If we first need to develop subtheories to develop high-scoring hybrid systems, we then lose the very point of the Newell Test.

Third, and most important, hybrid systems are artificially assembled systems and thus bear little true psychological and neurophysiological significance. Although we all agree that the human mind is a complex, multilevel construct and involves mechanisms and operations at, among others, both psychological and neuronal networks levels, simply piecing them together is ad hoc and trivializes the problem. A ToM that explains one phenomenon using a neural-network-level mechanism and explains another phenomenon using a rule-based, symbolic-level mechanism may be a convenient hybrid ToM, but is certainly not the *unified* ToM that Newell had wished for (cf. Newell 1990).

In our view, any principled ToM must recognize that the human mind may adopt different mechanisms and follow different laws at different levels. In addition, it is highly unlikely that there exists any simple and linear one-to-one mapping across levels. Penrose, for example, went so far as to hypothesize that there is a noncomputational and nonlocal process called "objective reduction" that connects physics and consciousness (Penrose 1996; see also Woolf & Hameroff 2001). We would argue that a similar nonlinear relationship exists between the neuronal-network-level and the psychological level, and that each level tells a veracious but adequately distinct story of mind. Such a multilevel view is also consistent with both Marr's (1982) and Newell's (1990) conception of multiple-level description of human cognition. Consequently, we should not expect a single architecture, even a hybrid one, to explain all of the phenomena of mind.

We regard both ACT-R and connectionism as celebratory candidates for a ToM, but at different levels. Whereas ACT-R focuses on the symbolic mental structures and processes and offers a psychologically plausible explanation that closely links to empirical behaviors, connectionism adopts subsymbolic neural-based mechanisms and permits a biologically realistic explanation

https://doi.org/10.1017/S0140525X03430136 Published online by Cambridge University Press

of mind that closely links to brain functions (e.g., O'Reilly & Munakata 2000). The two approaches are distinct in that symbols simply do not exist at a subsymbolic level. A unified ToM needs to encompass both levels of description, though each may be embodied in separate cognitive architectures. We regard the attempt to vertically stretch one level of analysis to linearly map to another as problematic. For example, we doubt that there is such a simple one-to-one mapping between ACT-R components and brain structures, as suggested in Figure 1 of the target article. It is hard to imagine (and not supported by neuroscience evidence) that the damage to the basal ganglia would completely destroy the work of mind, given the fundamental role that production rules play in ACT-R.

In summary, although we agree with Newell and A&L that a unified ToM is needed to advance cognitive science, we have different opinions regarding how to achieve such a unified theory. Our position is that, instead of hybridizing different approaches or linearly mapping them to boost the Newell Test score, we need to recognize the multilevel nature of the human mind and develop complementary theories at both psychological and connectionist levels, and cross-validate them.

NOTE

1. Hongbin Wang is the corresponding author for this commentary.

Newell's program, like Hilbert's, is dead; let's move on

Yingrui Yang[a] and Selmer Bringsjord[b]

[a]*Department of Cognitive Science;* [b]*Department of Computer Science, Rensselaer Polytechnic Institute, Troy, NY 12180.* **yangyri@rpi.edu selmer@rpi.edu http://www.rpi.edu/~brings**

Abstract: We draw an analogy between Hilbert's program (HP) for mathematics and Newell's program (NP) for cognitive modeling. The analogy reveals that NP, like HP before it, is fundamentally flawed. The only alternative is a program anchored by an admission that cognition is more than computation.

As most readers will know, Hilbert's program (HP) was devoted to building a system at or below the level of Turing machines (and their equivalents) to definitively settle all mathematical questions. Most readers will also know that in 1931, a young Viennese logician, Kurt Gödel, proved two incompleteness theorems – and Hilbert's program (HP) instantly died. Out of this death was born a much more sophisticated approach: In a word, true *meta* mathematics arose. One specific innovation was to devise and study infinitary logics not bound by Gödelian incompleteness, because, from an information-processing perspective, these logics are *beyond* Turing machines (Barwise 1980). Newell's program (NP) – the attempt to build a system at or below the level of Turing machines able to satisfy the criteria that Anderson & Lebiere (A&L) distill for us – has likewise expired. The difference is that apparently many influential cognitive scientists want to pretend the funeral never happened.

A&L, and in fact all those who see themselves as advancing NP, presuppose an exceedingly convenient sense of *universality.* According to this sense, a system is "universal" if and only if it can compute all Turing-computable functions. The construal is convenient because the vast majority of functions (in the relevant classes; e.g., functions from N to N) *aren't* Turing-computable (Boolos & Jeffrey 1989). A&L, the classical connectionists they evaluate, those they count as fans (e.g., Dennett), and so forth – all *assume* that human cognition can be nicely packaged beneath the Turing Limit. Yet, after decades of work, no system at or below the Turing Limit has the conversational power of a toddler (to pick just one criterion: 7). You would think the notion Newell (and his similarly sanguine partner, Simon) so confidently preached at the dawn of cognitive science (that thinking is computing at or below the Turing Limit, and computers operating at or below this limit with human-level intelligence will soon arrive) would, like Hilbert's dream, be a carcass at this point, but yet here is a *BBS* target article still stubbornly clinging to the religion (by cheerfully acting as if everyone is a believer). What arguments support the doctrine that cognition can be captured by standard computation? Surely no cogent argument is to be found on the basis of what has been built. LOGIC THEORIST, at the 1956 Dartmouth conference that kicked off AI, was able to prove the marvelously subtle theorem that *if p then q* implies *if not-q then not-p,* and this prompted Simon to declare that thinking machines would soon be among us. The situation is no different now: Here are A&L confidently pressing on to capture cognition in simple computation – but on the strength of what impressive artifact? Since seeing is believing, you will pardon us for not believing.

The problem isn't just criterion 7. Faced with consciousness, NP irremediably fails. Yet A&L once again cook up the convenient: They construe *consciousness* (in criterion 8) so that it simply leaves out the concept that threatens Newell's enterprise: namely, *phenomenal* consciousness. Block (1995) has recently explained the issue in this very journal. ACT-R and all forms of connectionism, and indeed every approach to sustaining NP, can't even take the first step toward expressing, in a third-person scheme, what it feels like to taste deep chocolate ice cream. ACT-R will be used to at most create what one of us (Bringsjord 2000) has called "zombanimals," that is, artificial animals with no inner lives. A robot with the behavioral repertoire of a dog, but with the inner life of a rock, might well be something NP, fueled by ACT-R, can produce.

That NP, as driven specifically by ACT-R, is dead, can be seen with help from concrete, not just philosophical, challenges. ACT-R is wholly incapable of modeling beliefs of the sort that human readers have when reading murder mysteries. For example, as detailed in Bringsjord (2000), readers have n-order beliefs about villains and detectives, and they make inferences based on these beliefs. For example, the reader of a murder mystery often believes that the villain believes that the detective believes that the villain believes the villain will never be caught. You can't represent this in ACT-R, period, because ACT-R is at best part of first-order logic devoid of doxastic operators. Of course, one could hook up a robust theory of human and machine reasoning (e.g., see Yang & Bringsjord, forthcoming) to ACT-R, but then in what sense *is* that new composite system ACT-R? If ACT-R is to be genuine science, it must be falsifiable. Yet A&L seem to describe an evolution in which serious challenges are handled by simply augmenting the system.

Just as the death of HP gave birth to infinitary logic, so should the death of NP give rise to cognitive modeling untrammeled by standard computation. Cognitive modelers need to step outside the notion that mere computation will suffice. They must face up to the fact, first, that the human mind encompasses not just the ordinary, humble computation that Newell and all his followers can't see beyond, but also *hyper*computation: information processing at a level *above* Turing machines, a level that can be formalized with help from analog chaotic neural nets, trial-and-error machines, Zeus machines, and the like (Bringsjord & Zenzen 2003; Siegelmann 1999).

In his famous "twenty questions" paper, Newell (1973) tells us that a sound science of the mind should not be steered by the willy-nilly dictates of experiment-driven empiricism. Instead, we are to do computational cognitive modeling. But such modeling, if limited by NP, fails to let cold hard reality lead the way. Instead, it lets speculative assumptions (e.g., that the mind processes information at or below the Turing Limit) prevent nature from proclaiming that we are more than ordinary machines.

https://doi.org/10.1017/S0140525X03430136 Published online by Cambridge University Press