The dilemma of unification and simplification in cognitive architectures

Coty Gonzalez

Dynamic Decision Making Lab
www.cmu.edu/ddmlab
Social and Decision Sciences Department
Carnegie Mellon University

Cognitive Architectures

- Unify: One aim of science
  - "… positing a single system of mechanisms – a cognitive architecture – that operate together to produce the full range of human cognition." (Newell, 1990)
  - Accumulate knowledge and applicability
  - Unification is served by the fact that the same set of basic processes is used to explain every cognitive phenomenon

- Simplify: Another aim of science
  - Simplicity = "Informativeness" = understandability = clarity = transparency
  - In the philosophy of science, simplicity is a criterion by which to evaluate competing theories
  - it becomes increasingly difficult to explain how ACT-R models work and how they are able to explain human behavior so well.

- **Can we Unify and simplify with cognitive architectures?**

Depart from the premise that: “All models are wrong”

- Cognitive architectures: big models that represent human cognition
- By definition, the ACT-R architecture is (Anderson et al., 2004):
  - Incomplete
  - Constrained
  - Not totally correct
  - Difficult to handle and to explain
- Representation of full human behavior is a very complex challenge.
- Some capabilities of cognitive architectures may (only) be attained through complex use of technology
  - Technological solutions that have nothing to do with the theory

Growing evidence for the need to simplify the technology

- Frank – already gave us many examples

- Bonnie John’s CogTool
- Dario Salvucci’s Distract-R
- Frank Ritter’s … many attempts

- Coty Gonzalez’s IBLTool

Simplification should come not only in the software

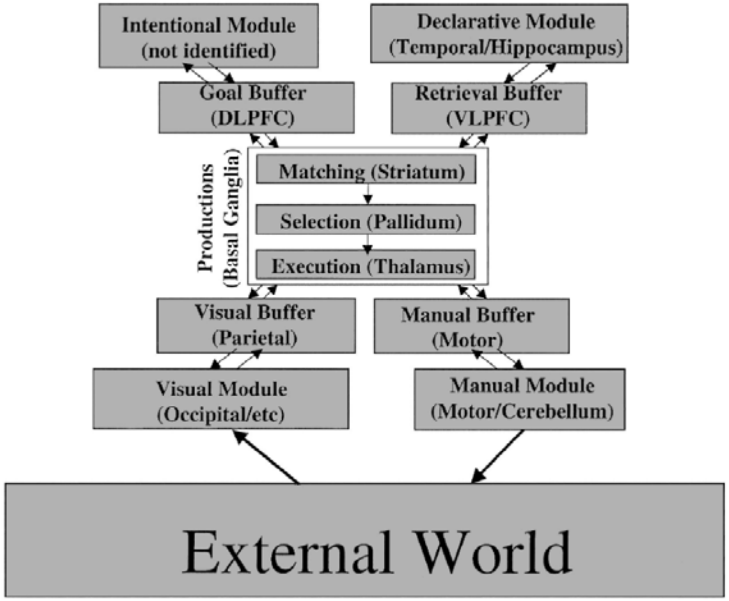
- Transparency of the mechanisms
  - Matlab
  - Excel
  - Having to explain the whole ACT-R theory in papers

- Scientific progress: going deeper rather than wider

- Test and Validation of theories
  - Model comparison

1. Possible solutions to the dilemma: ACT-R as a toolbox

- The modular yet integrated view of ACT-R is a good idea:
  - Great conceptual illustration of the modules involved in the cognitive system
  - Mapping to cortical regions
- BUT:
  - What exactly does each “tool” in the box do? – what is a tool?
  - What are exactly the practical implications of “Buffers” in the representation of human behavior?
  - Why do we call something a “module”?
  - How are tools recruited? When is each tool needed? Do we really need to “put it all together”?
  - How do the tools interact?


Intentional Module (not identified)
Declarative Module (Temporal/Hippocampus)
Goal Buffer (DLPFC)
Retrieval Buffer (VLPFC)
Productions (Basal Ganglia)
Matching (Striatum)
Selection (Pallidum)
Execution (Thalamus)
Visual Buffer (Parietal)
Manual Buffer (Motor)
Visual Module (Occipital/etc)
Manual Module (Motor/Cerebellum)
External World

ACT-R as a toolbox

- The “tools” are not the modules, but the set of equations and parameters
- Unpack the equations
  - Determine when and how and why each component of each equation is needed. For example,
    - IBLT, Gonzalez, Lerch, Lebiere, 2003: full activation equation, blending, similarity
- Repeated binary-choice and sampling tasks (Lebiere, Gonzalez, & Martin, 2007; Lejarraga et al., 2010; Gonzalez & Dutt, 2011)
  - Technion Prediction Competition (Erev et al., 2010)
  - Visual basic implementation of the IBL ideas
  - Matlab and Excel implementation of IBL model for repeated choice tasks
  - More generic IBLTool

Learning mechanisms for repeated choice

Gonzalez & Dutt, 2011; Dutt & Gonzalez, 2011; Lejarraga, Dutt & Gonzalez, 2010; Lejarraga, Dutt & Gonzalez, in prep

- Choose the option with the highest “blended” value :

$$V_j = \sum_{i=1}^{n} p_i x_i$$

- The probability of retrieval is a function of memory Activation (A) of that outcome relative to the activation of all the observed outcomes for that option given by:

$$P_{i,t} = \frac{e^{A_{i,t}/\tau}}{\sum_j e^{A_{j,t}/\tau}} \qquad \tau = \sigma \cdot \sqrt{2}$$

- Activation: simplification of ACT-R’s mechanism (Anderson & Lebiere, 1998):

$$A_{i,t} = ln\left(\sum_{t_i \in \{1,\ldots,t-1\}} (t - t_i)^{-d}\right) + \sigma \,.\, ln\left(\frac{1-\gamma_{i,t}}{\gamma_{i,t}}\right)$$

- 2 free parameters:
  - *Noise:* σ : high s -> high variability if retrieval
  - *Decay: d* : high d-> More recency

ACT-R as a toolbox

- Unpack the parameters
  - What do the parameters mean?; What do the default values mean?
    - Individual differences work
    - Task and environmental differences

2. Possible solutions to the dilemma: Unified mini-theories

- Theories that use a subset of ACT-R mechanisms.
  - Salvucci & Taatgen's: A unified theory of multi-tasking
  - Gonzalez, Lebiere and others: Instance-Based Learning Theory
- Constrain the current freedom of an ACT-R modeler:
  - Freedom in approaches to develop a representation of the behavior for a task.
  - Freedom in choosing equations and parameter values in order to "fit" model data to human data.

Conclusions

- Allen Newell (1990), cognitive architectures goal of unification presents a dilemma to simplification
  - Simplification is not well served in cognitive architectures
  - It has been neglected in ACT-R

- Possible ways to deal with the dilemma:
  - ACT-R as a toolbox where the tools are the equations and parameters: Transparency and validation of equations and parameters
  - Unified “mini-theories”
  - The creation of explicit computer tools that represent a “mini-theory” can give rise to interesting demonstrations and new questions and answers