Metrics for Cognitive Architecture Evaluation

Bob Wray & Christian Lebiere
wray@soartech.com
23 May 2007

Evaluation of Soar in UTC (Functional)

Behave flexibly	yes
Exhibit adaptive behavior	yes
Operate in real-time	yes
Operate in rich environment • Perceive dynamic details • Use vast amounts of knowledge • Control a motor system	 interface (yes) yes interface (yes)
Use symbols and abstractions	yes
Use language	no (yes)
Acquire capabilities via development	no
Operate within a social community	no (yes)
Be self-aware	no

- "Yes" indicates a demonstration of satisfaction of these requirements?
- In some particular case?
- To what extent?

Evaluation of Soar in AIJ (Coverage)

Knowledge-lean tasks • AI toy problems (e.g. blocks world)
Small routine tasks • Unification, syllogisms, etc.
Knowledge-intensive expert-system tasks: • R1-Soar, NeoMYCIN, DESIGNER
Miscellaneous AI Tasks/Capabilities: • Language processing, *planning*, etc.
Learning • Learns on all tasks it performs (whoops…) • Practice, transfer, strategy acquisition, operator implementation, macros, EBL • *Conceptual, instruction & observation, error recovery, reinforcement, …*
Dynamic, knowledge-based control & interaction • *Robotic control (Hero-Soar, Air-Soar)* • *Human Behavior Representations (TacAir-Soar)* • *Multi-agent Systems (IDA)*

- "Box scores don't belong is science…. But some way is needed to emphasize how important coverage is." (UTC)

"Newell Test" Evaluations of ACT-R & Classical Connectionism

Functional Criteria	ACT-R	CC	Soar
Behave flexibly	*Better*	*Mixed*	
Exhibit adaptive behavior	*Better*	*Better*	
Operate in real-time	*Best*	*Worse*	
Vast knowledge of rich environments	*Mixed*	*Worse*	
Knowledge integration/distal access	*Mixed*	*Worse*	
Use language	*Worse*	*Better*	
Learn from environment	*Better*	*Better*	
Be self-aware	*Worse*	*Worse*	

- "Grades" are comparative within each theory; not across theories.
- Adapted from Anderson, J. R., & Lebiere, C. (2003). The Newell test for a theory of cognition. *Behavioral and Brain Science, 26*, 587-637

Why we need better evaluations…

- Newell's criteria are not solidly defined:
  - "Operate in real time"
  - "Adaptive behavior"
- Instance-based grading is not really that insightful
  - One application/model is insufficient for "grading"
  - Anderson & Lebiere survey approach is still "soft"/subjective
- Generality is the fundamental goal of cognitive architectures
  - Generality implies a significant utility/value over many different types of problems
  - Benchmark performance is the current gold-standard in AI
  - Application-specific solutions almost always "out perform" cognitive-architecture-based solutions in CPU/thruput/etc.
  - Critical need in "applied cognitive architecture" is ways to make the power of cognitive architecture approach evident in a metrics-driven funding environment
- How can we evaluate and communicate progress toward generality in the language of science (empirical demonstrations of phenomena)?

Can we improve on evaluations?

- Start taking each Newell-Test criteria seriously: What measures could be applied?
- Understand and present performance measures in the context of the Newell-Test criteria

Creating Metrics

- Newell Test outlines what we want to measure
  - *How* do we measure the desired characteristics?
- Solutions/directions:

  - Objective, problem-independent measures
    - Cognitive operations, response time, incrementality
  - Problem/solution-specific quantitative metrics
    - Adaptivity
  - Subjective consensus rankings
    - Problem complexity judgments
  - Enumerations ("Box scores")
    - Versatility

- Goal here is to ***begin*** the process of defining good (objective, quantitative) metrics for many cognitive-architecture-based applications
  - Appropriate metrics enable hypothesis-driven scientific exploration
    - Tools for asking questions
  - Specific problems will drive which metrics become more fully elaborated

Newell Test Criterion

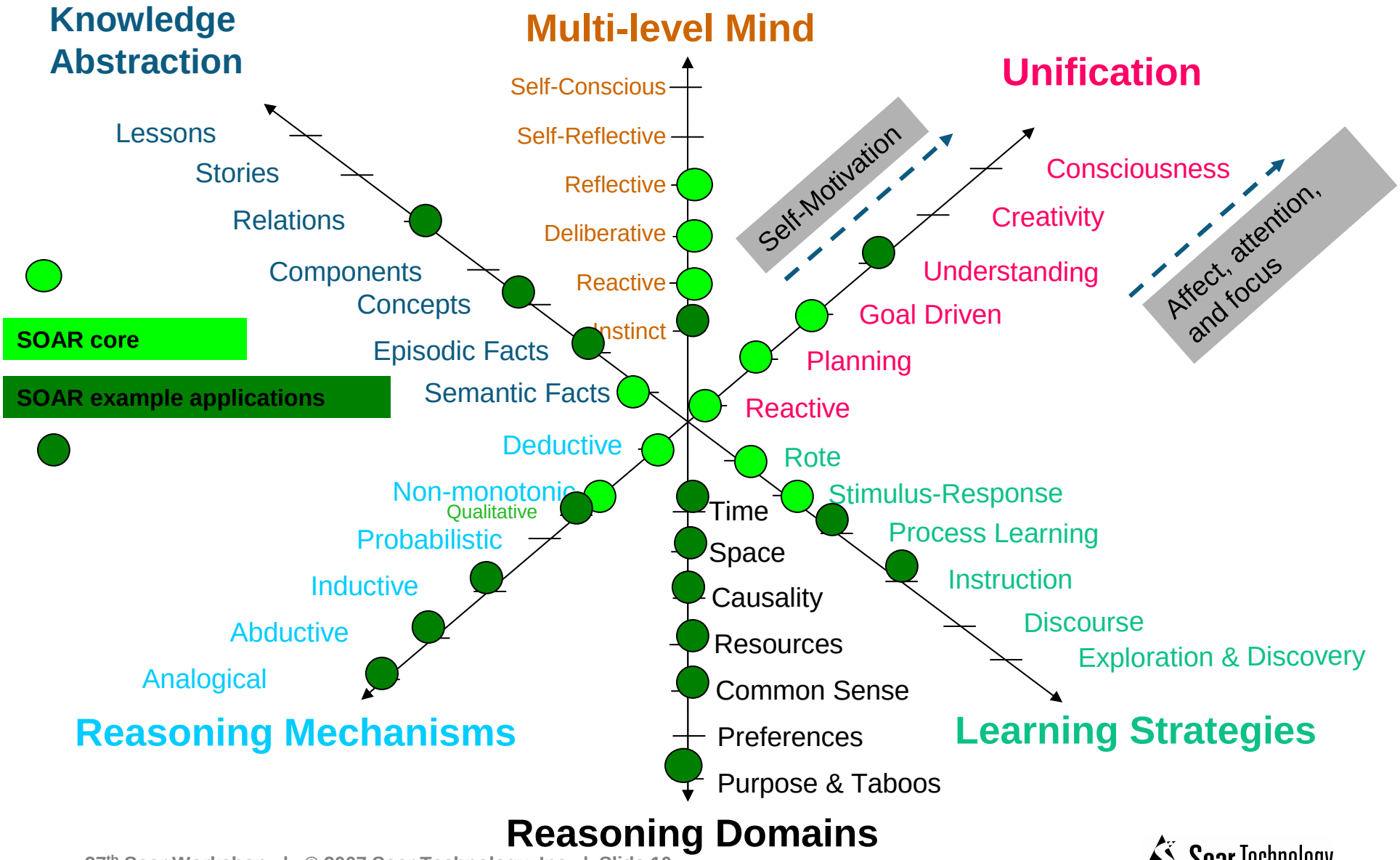
- Anderson & Lebiere
  - How do Anderson and Lebiere define this criterion? How do they "grade" it?
- Evaluation focus:
  - How might we define the criterion in fully functional terms?
- Measures
  - Ideas for measures for this criterion

Behave flexibly

- Anderson & Lebiere:
  - Universal computation
  - “ability to learn to perform almost arbitrary cognitive tasks to high degrees of expertise…. [with] no anticipation in human evolutionary history.”
- Evaluation focus:
  - Breadth and autonomy of capability
- Measures:
  - Versatility: Box score list of domains
  - Taskability
    - Domain-specific measures of transfer and novelty (e.g., Transfer Learning?)

from Soar 23 presentation

Enumeration: Versatility



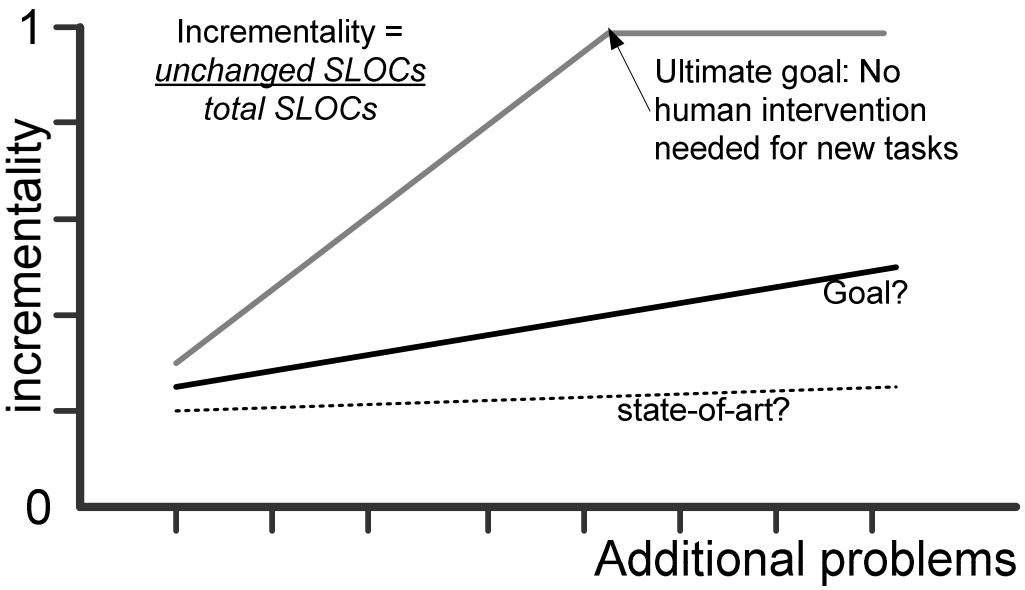
Soar Technology
Thinking *inside* the box.

Behave flexibly

- Anderson & Lebiere:
  - Universal computation
  - “ability to learn to perform almost arbitrary cognitive tasks to high degrees of expertise…. [with] no anticipation in human evolutionary history.”
- Evaluation focus:
  - Breadth and autonomy of capability
- Measures:
  - **Versatility**: Box score list of domains
  - **Taskability**
    - Domain-specific measures of transfer and novelty (e.g., Transfer Learning?)
  - **Incrementality**
    - Measure reuse from one application to the next

Measure: Incrementality

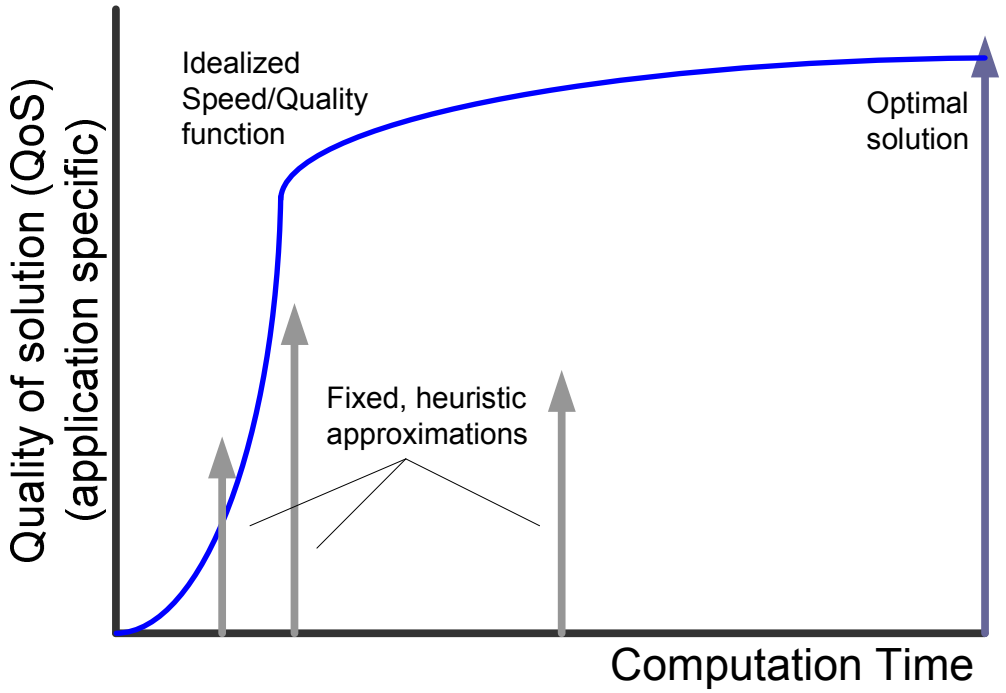
- Simple measure
  - Analog to SE?
  - Simple tools (diff?)
  - Decompose for more fine-grained comparisons
    - Soar 7 vs. 8 . vs. 9
    - TAS vs. MOUTBots



- Value
  - Immediate transparency of reuse from one application to another
  - Motivator for reuse of knowledge-representation-level capabilities?
  - Driver for general, usable “default rules” capabilities: planning, impasse resolution, etc.

Operate in Real-time

- Anderson & Lebiere:
  - How can cognition be simulated in “human time”
- Evaluation focus:
  - Measures of actual performance
  - Be explicit about cost of analytic solutions, lack of anytime properties
  - Translate to human simulation time for cognitive models



Typical Performance Metrics

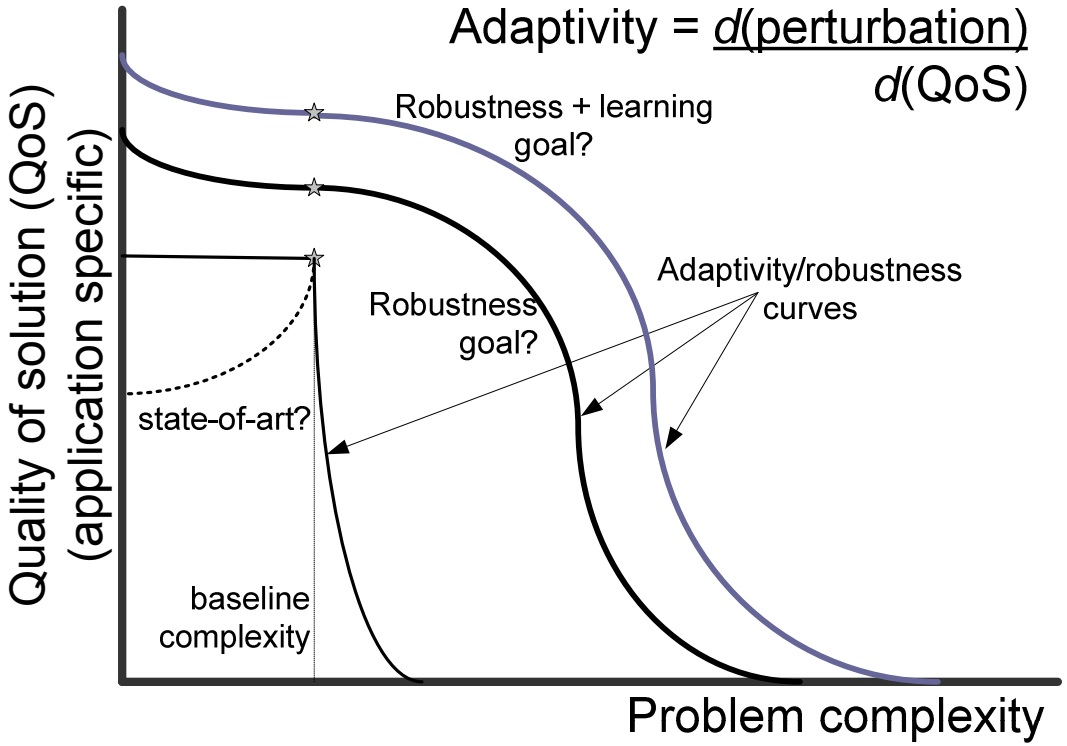
- Cognitive operations/unit time
  - Throughput of the cognitive architectures (decisions/sec)
- Response time
  - Time to respond to a particular problem or situation
- Footprint
  - Memory, CPU %, interconnect bandwidth, etc.
- All performance measures are *relative:*
  - to the problem
  - to the application/implementation
  - to the hardware implementation
  - Many (most?) do not understand the relativity of performance measures!

Exhibit adaptive behavior

- Anderson & Lebiere:
  - Rational analysis; architectural and system-level adaptation to actual environment
- Evaluation focus:
  - How well does a system behave in situations it was not specifically designed for?
- Measures:
  - **Adaptivity**: Measure how well a system responds to perturbations in the task environment
  - Performance measures:
    - E.g., How does response time change in under perturbation conditions?

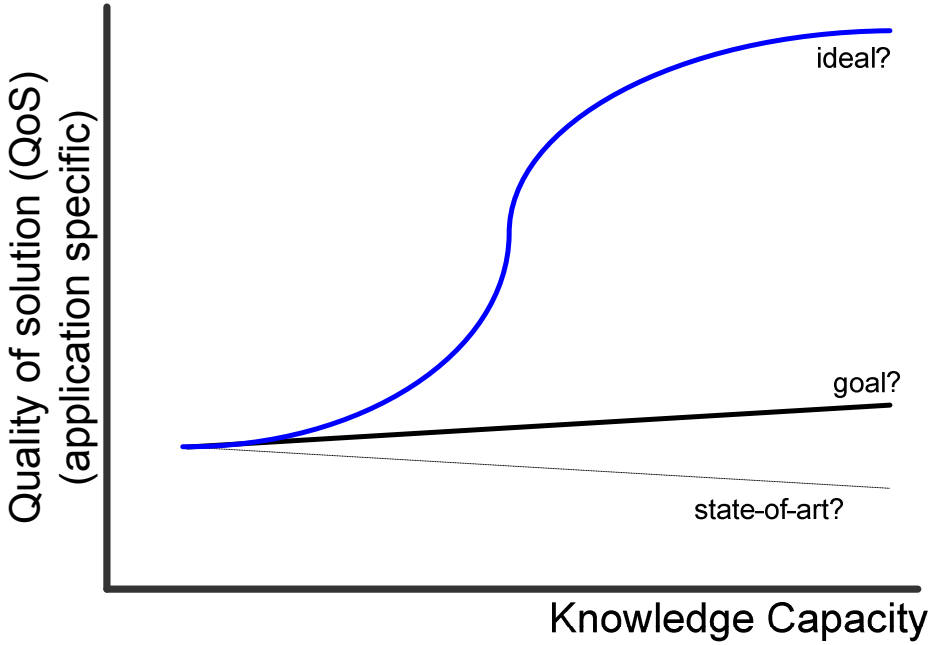
Measure: Adaptivity

- Response of system to perturbation in the "designed" operational environment
- Not a learning metric, but adaptivity may be improved by learning
- Analog to stability region in control systems
- Limitations:
  - Domain-/problem-specific
  - Requires "problem complexity" dimension
  - Requires baseline performance system



Use vast amounts of knowledge

- Anderson & Lebiere:
  - Determine how performance changes with the scale of the knowledge base
- Evaluation focus:
  - Knowledge scalability
- Measures:
  - **Knowledge capacity**: How much "knowledge" is encoded in an agent?
  - **Knowledge utilization**: How much of the encoded "knowledge" is actually used in solving a suite of problems?
  - Performance measures:
    - How do performance measures (COPS, footprint, response time) change with knowledge capacity and knowledge utilization?



Behave Robustly

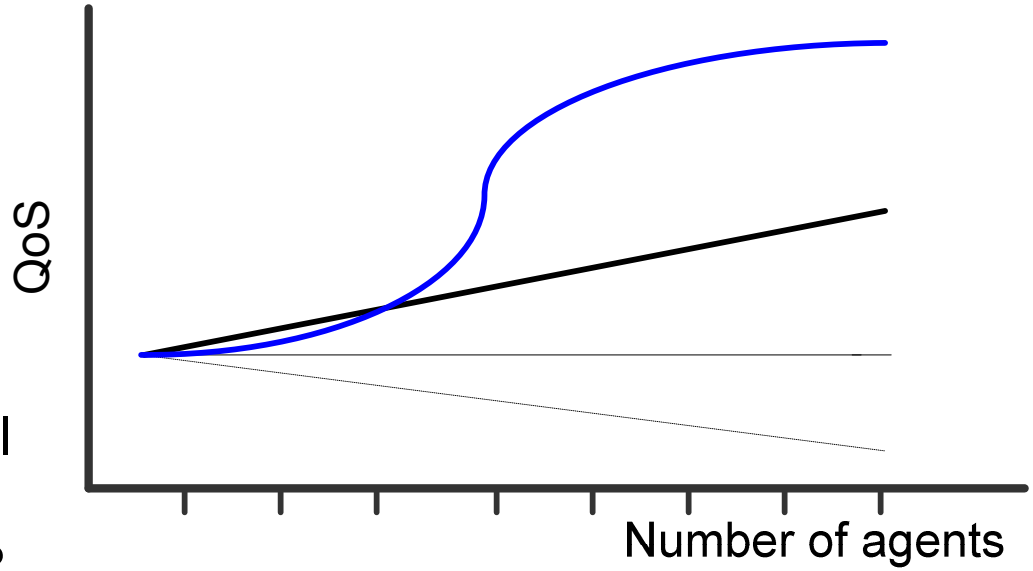
- Not included in Anderson & Lebiere list (likely included in their view of adaptivity)
- Evaluation focus:
  - Ability to handle uncertain, incomplete, stochastic information
  - Different from adaptivity (but related)
    - Highlights ability to cope with uncertain and incomplete information within the basic task performance space
- Measures:
  - **Robustness**: Domain-specific measures (e.g., Nielsen, Beard et al, 2002)
  - **Stochastic assimilation**: Ability to capture and express stochastic distributions in the environment

Integrate knowledge

- Anderson & Lebiere:
  - Produce intellectual activities that are the hallmarks of human capacity for intellectual combination…. things like inference, induction, metaphor, and analogy.
- Evaluation focus:
  - Versatility: Enumeration of capabilities (for now?)
- Open questions:
  - How can we avoid/diminish “wishful thinking”?

Behave autonomously in a social environment

- Anderson & Lebiere
  - Mostly focus on language, rather than the social environment generally
- Evaluation focus:
  - To what extent does social environment impact the performance of the agent?



- Measures:
  - **Scalability**: How do increasing numbers of agents impact the quality of solution?

Learn from the environment

- Anderson & Lebiere:
  - Demonstrations of learning across Squire's (1992) taxonomy of human memory and learning
- Evaluation focus:
  - Functional impacts of learning
- Measures:
  - Performance measures (faster performance with learning)
  - Changes in knowledge capacity and knowledge utilization due to learning
  - Domain-specific measures to demonstrate qualitative changes in capability

Exhibit self-awareness & sense of self

- Anderson & Lebiere:
  - Focus on implicit learning
- Evaluation focus:
  - Enumeration of functional roles of “consciousness”
- Measures:
  - Adaptivity and robustness may include elements of self-awareness?

Conclusions

- Coal
  - Community currently lacks convincing tools to demonstrate (empirically, scientifically) the (assumed) value of cognitive architectures
  - Lack of empirical demonstrations to substantiate claims is a serious hole in the cognitive-architecture narrative
  - Metric definition is hard, especially for complex domains
- Gold
  - Functional emphasis of Soar is beneficial in helping define metrics
  - There is some low-hanging fruit (incrementality, knowledge capacity)
  - We can do better… and communicate empirical results