From Hybrid Systems to Hybrid Cognitive Architectures

Ron Sun

Background:

— From hybrid systems to hybrid cognitive architectures

combining a variety of techniques
possessing a variety of capabilities/functionalities
more comprehensive systems

⟶ domain-generic models of cognition

— From engineering to science

empirical scientific data
empirical validation
theoretically driven

hypothesis-test

— A difficult problem and a grand challenge

What is a Cognitive Architecture?

A cognitive architecture is a broadly-scoped, domain-generic computational cognitive model, capturing the essential structure and process of the mind, to be used for a broad, multiple-level, multiple-domain analysis of cognition and behavior; Sun (2004, *Philosophical Psychology*)

— Architecture of a building:

overall framework and overall design, as well as roofs, foundations, walls, windows, floors, and so on.

— Cognitive architecture:

overall structures, essential divisions of modules, relations between modules, basic representations, essential algorithms, and a variety of other aspects.

— Relatively invariant across time, domains, and individuals

— Structurally and mechanistically well defined

— Componential processes of cognition

Functions

in relation to cognitive science and in relation to artificial intelligence:

— To provide an essential framework to facilitate more detailed modeling and understanding of various components and processes of the mind

specifying computational models of cognitive mechanisms and processes

embodying descriptions of cognition in computer programs

— To provide the underlying infrastructure for building intelligent systems

including a variety of capabilities, modules, and subsystems

implementing understanding of intelligence gained from studying the human mind

⟶ so, more cognitively grounded intelligent systems

Why are Cognitive Architectures Important for Cognitive Science?

— Psychologically oriented cognitive architectures: “intelligent” systems that are cognitively realistic, cognitive theories that have been validated through psychological data, and so on

— They shed new light on human cognition and therefore they are useful tools for advancing the science of cognition

— They may even serve as a foundation for understanding collective human behavior and social phenomena

Why are Cognitive Architectures Important for Cognitive Science?

— Force one to think in terms of process, that is, in terms of mechanistic (computational) detail

— Require that important elements of a model be spelled out explicitly, thus leading to conceptually clearer theories

— Provide a deeper level of explanation, not centered on superficial, high-level features of a task

— Lead to unified explanations for a large variety of cognitive data and/or cognitive phenomena

— Develop generic models of cognition capable of a wide range of cognitive functionalities, to avoid the myopia of narrowly-scoped research

Newell (1990), Sun (2002, book published by Erlbaum)

— In all, cognitive architectures are believed to be essential in advancing understanding of the mind (Anderson 1983, Newell 1990, Anderson and Lebiere 1998, Sun 2002)

— Therefore, developing cognitive architectures is an extremely important enterprise in cognitive science

Why are Cognitive Archtiectures Important for AI/CI?

— Support the central goal of AI/CI: building artificial systems that are as capable as human beings

— Help us to reverse engineer the only truly intelligent system around—the human being

— Form solid basis for building truly intelligent systems, because they are well motivated by, and properly grounded in, existing cognitive research

— Facilitate the interaction between humans and artificially intelligent systems

— Antithesis of expert systems: instead of focusing on capturing performance in narrow domains, they are aimed to provide broad coverage of a wide variety of domains

— Many current business/industrial applications of intelligent systems increasingly require broad systems that exhibit a broad range of intelligent behaviors, not just isolated systems of individual functionalities

For example, one application may require the inclusion of capabilities for raw image processing, pattern recognition, categorization, reasoning, decision making, and natural language communications. It may even require planning, monitoring, control of robotic devices, and interactions with other systems and devices

— Such requirements highlight the importance of research efforts on broadly scoped cognitive architectures that perform a broad range of cognitive functionalities across a variety of task domains

Still Room for Grand Theories?

— Some have claimed that fundamental scientific discovery and grand scientific theorizing have become a thing of the past. What remains to be done is filling in details and refining some relatively minor points

— Researchers in cognitive science are pursuing integrative approaches that explain data in multiple levels, domains, and functionalities

— Significant advances may be made through hypothesizing and confirming deep-level principles that unify superficial explanations across multiple domains

— Cognitive architectures are the basis of such unified theories (see, e.g., Sun 2002, the Erlbaum book)

An Example of a Cognitive Architecture

CLARION

— An integrative architecture, consisting of a number of distinct subsystems

— A dual representational structure in each subsystem (implicit versus explicit representations)

— Its subsystems include:

the action-centered subsystem (the ACS), the non-action-centered subsystem (the NACS), the motivational subsystem (the MS), and the meta-cognitive subsystem (the MCS)

their respective roles

— See Sun (2002, the Erlbaum book) and Sun (2003, Techinical specification)

Overview of CLARION

— Each subsystem consists of two levels of representation

that is, a dual representational structure

— The top level encodes explicit knowledge

— The bottom level encodes implicit knowledge

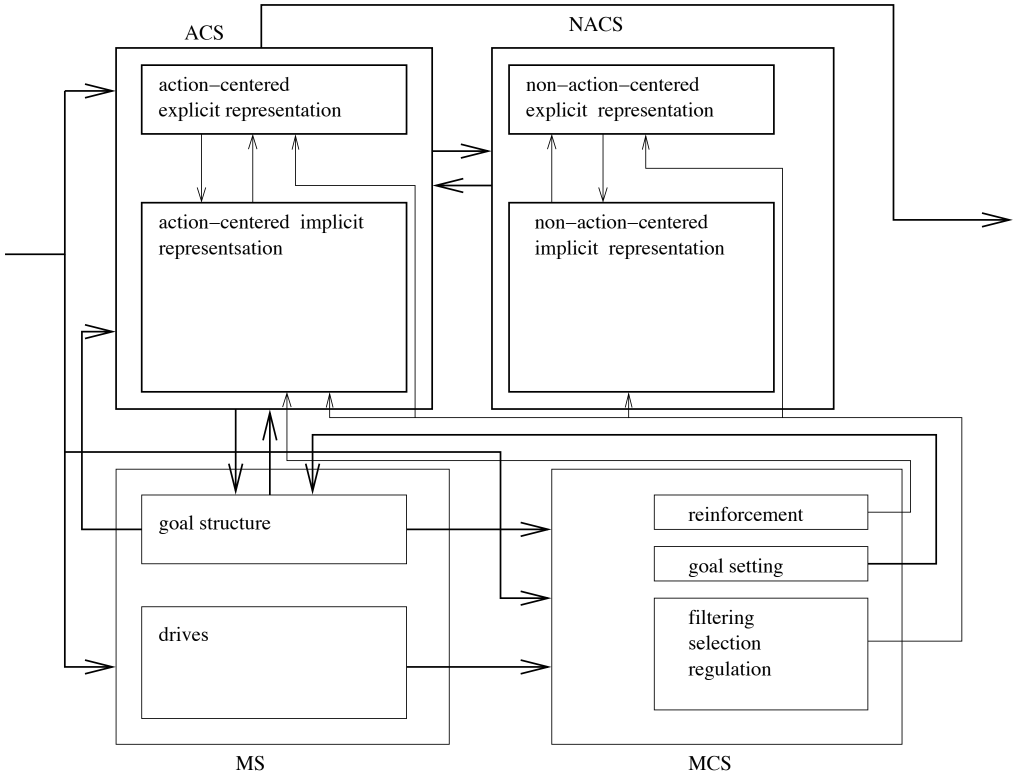
Reber (1989), Seger (1994), Cleeremans et al (1998), Sun (2002)

— The two levels interact, for example, by cooperating in actions

— Essentially, it is a dual-process theory of mind (Chaiken and Trope 1999)

— Duality of representation: extensively argued in Sun et al (2005; in *Psychological Review*)

Overview of CLARION



Some Details

The Action-Centered Subsystem:

The operation of the action-centered subsystem:

1. Observe the current state x.
2. Compute in the bottom level the "values" of x associated with each of all the possible actions a_i's: $Q(x, a_1)$, $Q(x, a_2)$, ..., $Q(x, a_n)$, based on implicit knowledge.
3. Find out all the possible actions (b_1, b_2, ..., b_m) at the top level, based on the input x (sent up from the bottom level) and the explicit knowledge (explicit rules) in place.
4. Compare or combine the values of the selected a_i's with those of b_j's (sent down from the top level), and choose an appropriate action b.
5. Perform the action b, and observe the next state y and (possibly) the reinforcement r.
6. Update Q-values at the bottom level in accordance with the *Q-Learning-Backpropagation* algorithm
7. Update the rule network at the top level using the *Rule-Extraction-Refinement* algorithm.
8. Go back to Step 1.

The Action-Centered Subsystem:

— In the bottom level of the action-centered subsystem, implicit reactive routines are learned:

Q-value

Q-learning

modularity

— In the top level of the action-centered subsystem, explicit conceptual knowledge is captured in the form of explicit symbolic rules

— See Sun et al (2001, *Cognitive Science*) and Sun (2003, Technical Specification) for details

The Action-Centered Subsystem: Autonomous generation of explicit conceptual structures

— The basic process of bottom-up learning:

If an action implicitly decided by the bottom level is successful, then the agent extracts an explicit rule that corresponds to the action selected by the bottom level and adds the rule to the top level. Then, in subsequent interaction with the world, the agent verifies the extracted rule by considering the outcome of applying the rule: if the outcome is not successful, then the rule should be made more specific and exclusive of the current case; if the outcome is successful, the agent may try to generalize the rule to make it more universal.

— A kind of rational reconstruction of implicit knowledge

— Learning explicit conceptual representation at the top level can also be useful in enhancing learning of implicit reactive routines at the bottom level

Sun et al (2001), Sun et al (2005)

— After explicit rules have been learned, a variety of explicit reasoning methods may be used

Sun (2003)

The Action-Centered Subsystem: Assimilation of externally given conceptual structures

— CLARION can learn even when no a priori or externally provided knowledge is available

— However, it can make use of it when such knowledge is available

— Externally provided knowledge, in the forms of explicit conceptual structures (such as rules, plans, categories, and so on), can

(1) be combined with existent conceptual structures at the top level

(2) be assimilated into implicit reactive routines at the bottom level

— This process is known as top-down learning

The Non-Action-Centered Subsystem

— Representing general knowledge about the world

that is, the “semantic” memory (Quillian 1968)

— Performing various kinds of memory retrievals and inferences

— Under the control of the action-centered subsystem (through its actions)

The Non-Action-Centered Subsystem

— At the bottom level, “associative memory” networks encode non-action-centered implicit knowledge

Distributed representation of microfeatures

— At the top level, a general knowledge store encodes explicit non-action-centered knowledge

Symbolic/localist representation of concepts, i.e., chunks (linked to microfeatures)

A node is set up in the top level to represent a chunk (a concept), and connects to its corresponding microfeatures (distributed rpresentation) in the bottom level

— At the top level, links between chunks encode associations between pairs of chunks (concepts), known as associative rules

The Non-Action-Centered Subsystem

— Similarity-based reasoning may be employed

During reasoning, a known (given or inferred) chunk may be automatically compared with another chunk. If the similarity between them is sufficiently high, then the latter chunk is inferred.

— Mixed rule-based and similarity-based reasoning

Accounting for a large variety of commonsense reasoning patterns (including "inheritance reasoning"); see Sun (1994, book published by Wiley), and Sun (1995, *Artificial Intelligence*)

The Non-Action-Centered Subsystem

— Top-down learning in the NACS

— Bottom-up learning in the NACS

The Motivational Subsystem

— Drives and their interactions (Toates 1986) leads to actions

It is concerned with why an agent does what it does. Simply saying that an agent chooses actions to maximizes gains, rewards, reinforcements, or payoffs leaves open the question of what determines these things

— It provides the context in which the goal and the reinforcement of the action-centered subsystem are set

— A bipartite (dual representational) system of motivational representation:

The explicit goals of an agent
may be generated based on internal drive states

The Motivational Subsystem

— Low-level primary drives (mostly physiological): hunger, thirst, danger, ...

— High-level primary drives (mostly social): seeking of social approval, desire for reciprocation, interest in exploration, ...

— Secondary (derived) drives

While primary drives are built-in and relatively unalterable, there are also "derived" drives, which are secondary, changeable, and acquired mostly in the process of satisfying primary drives

Derived drives may include: (1) gradually acquired drives, through "conditioning"; (2) externally set drives, through externally given instructions

The Motivational Subsystem

— A generalized notion of “drive”

— Essential desiderata (Tyrell 1993, Toates 1986, Hull 1943, Sun 2003)

— Activation levels of drives: determined by equations derived from essential desiderata (Sun 2003)

For example,

Get-food: *0.95 ** max *(food-deficit, food-deficit * food-stimulus)*

Avoid-danger: *0.98 * danger-stimulus * danger-certainty*

The Meta-Cognitive Subsystem

— Meta-cognition refers to "one's knowledge concerning one's own cognitive processes and products" and the control and regulation of them (Flavell 1976)

— Regulates not only goal structures but also cognitive processes per se

Schwartz and Shapiro (1986), Metcalfe and Shimamura (1994), Reder (1996), Mazzoni and Nelson (1998),

The Meta-Cognitive Subsystem

(1) behavioral aiming:
setting of reinforcement functions
setting of goals

(2) information filtering:
focusing of input dimensions in the ACS
focusing of input dimensions in the NACS
selection of input values (within certain input dimensions) in the ACS
selection of input values (within certain input dimensions) in the NACS

(3) information acquisition:
selection of learning methods in the ACS
selection of learning methods in the NACS

(4) information utilization:
selection of reasoning methods in the top level of the ACS
selection of reasoning methods in the top level of the NACS

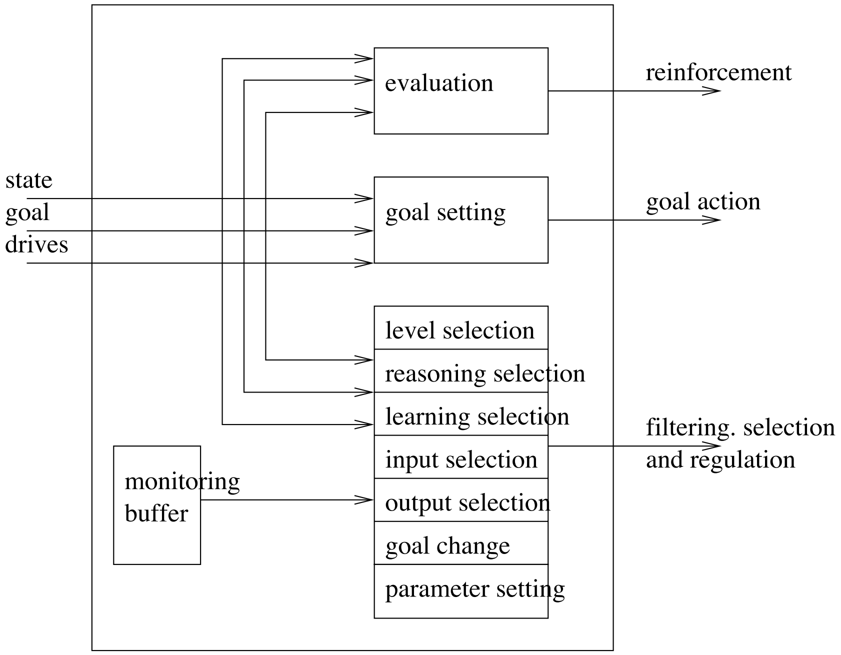
(5) outcome selection:
selection of output dimensions in the ACS
selection of output dimensions in the NACS
selection of output values (within certain output dimensions) in the ACS

selection of output values (within certain output dimensions) in the NACS

(6) cognitive modes:
selection of explicit processing, implicit processing, or a combination thereof (with proper integration parameters), in the ACS
selection of explicit processing, implicit processing, or a combination thereof (with proper integration parameters), in the NACS

(7) parameters of the ACS and the NACS:
setting of parameters for the IDNs
setting of parameters for the ARS
setting of parameters for the AMNs
setting of parameters for the GKS

Meta-Cognitive Subsystem: Structure



Structure of the meta-cognitive subsystem.

Accounting for Cognitive Data: Past simulations using CLARION

— Process control tasks

Berry and Broadbent (1988),
Stanley et al. (1989),

Dienes and Fahey (1995),

— Serial reaction time tasks

Lewicki et al. (1987),

Curran and Keele (1993)

— Artificial grammar learning tasks

Domangue et al (2004)

— Alphabetic arithmetic (letter counting) tasks

Rabinowitz and Goldberg (1995)

— Categorical inference tasks

Sloman (1998)

— Discovery tasks

Bowers et al (1986)

— Tower of Hanoi

Gagne and Smith (1962)

— Minefield navigation

Sun et al. (2001)

— “Lack of knowledge” inferences

Gentner and Collins (1991)

— Meta-cognitive monitoring

Metcalfe (1986)

focus: capturing the interaction, and the resulting synergy
using mainly bottom-up learning

Accounting for Cognitive Data

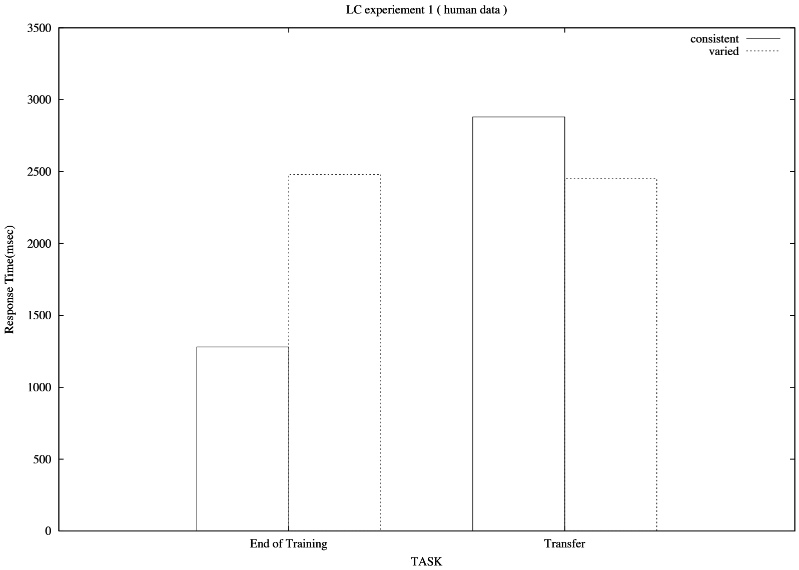
Two examples:

— Alphabetic arithmetic

— Process control

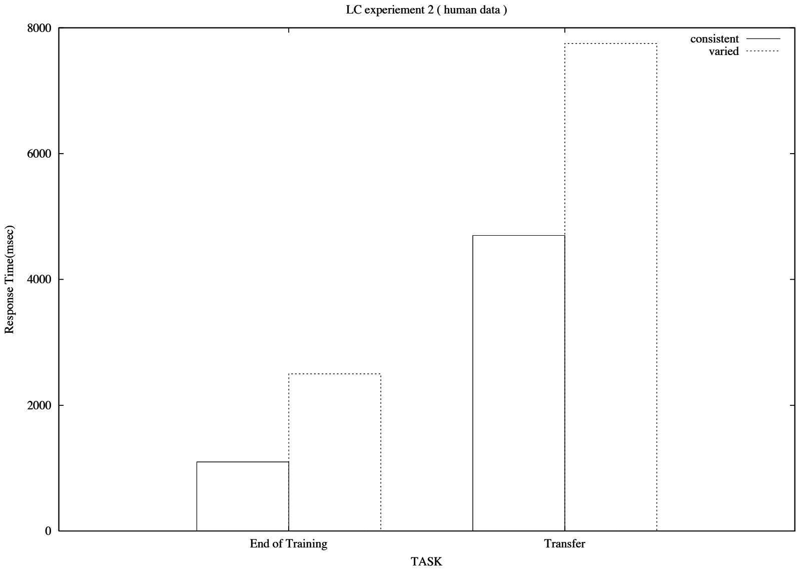
The Letter Counting Task: Rabinowitz and Goldberg (1995)

Experiment 1: $letter1{+}number = letter2$ or $letter1 - number = letter2$
— the consistent group: 36 blocks of training (the same 12 addition problems in each)
— the varied group: 6 blocks of training (the same 72 addition problems in each)
— transfer: 12 new addition problems (repeated 3 times)

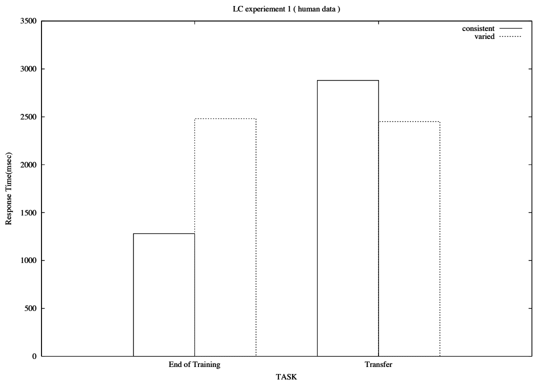


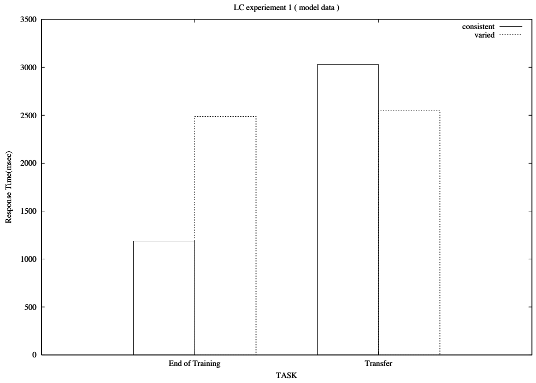
The Letter Counting Task: Rabinowitz and Goldberg (1995)

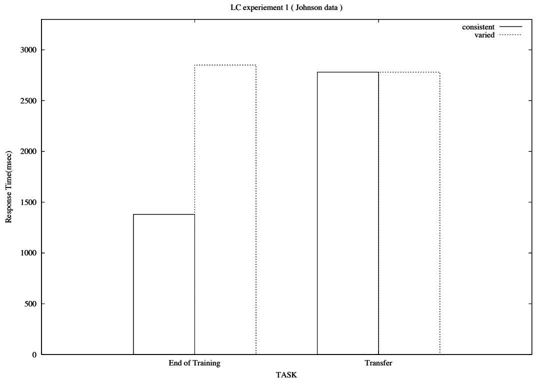
Experiment 2:
— training: the same
— transfer: 12 subtraction problems (repeated 3 times)



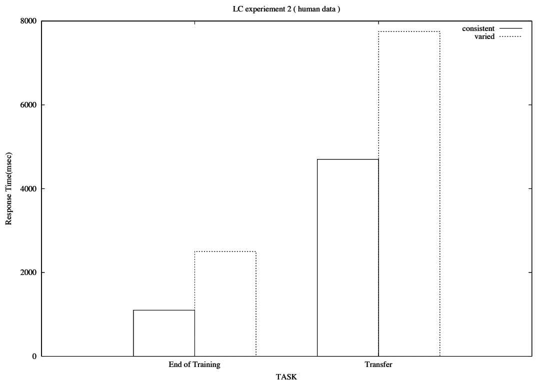
Simulation of experiment 1: Clarion vs. ACT-R

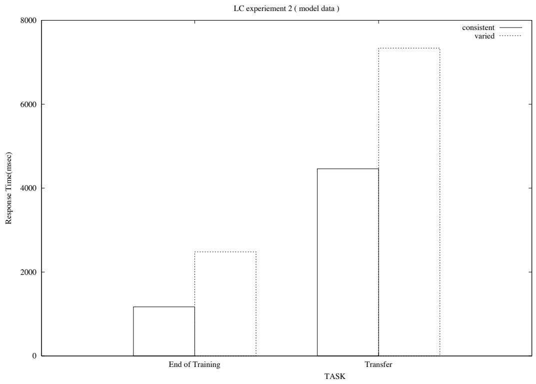


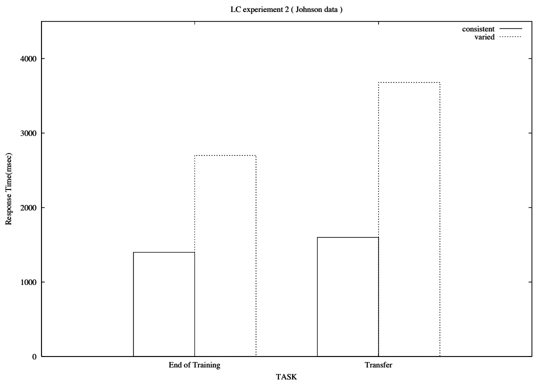




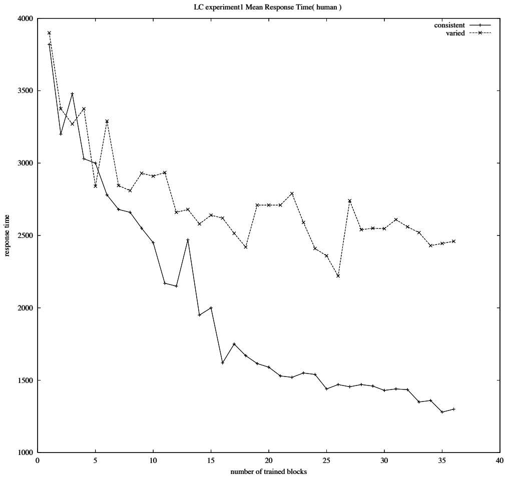
Simulation of experiment 2: Clarion vs. ACT-R



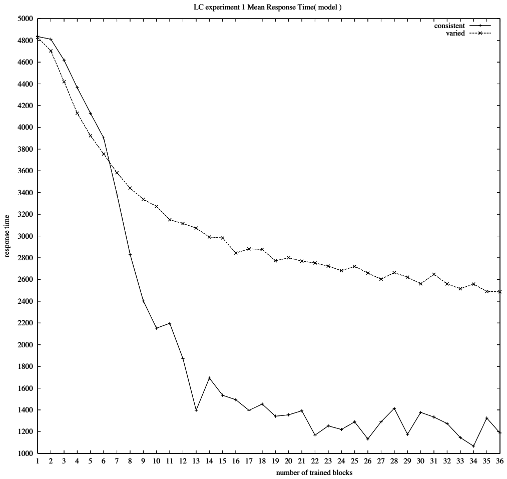




Simulation: learning curves

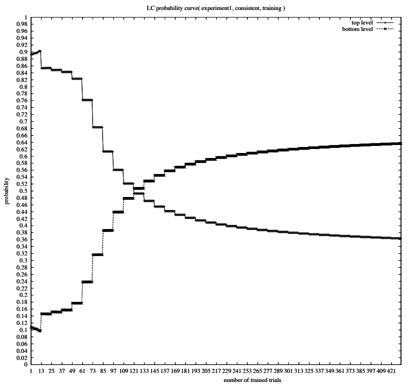


The learning curve of Rabinowitz and Goldberg (1995).

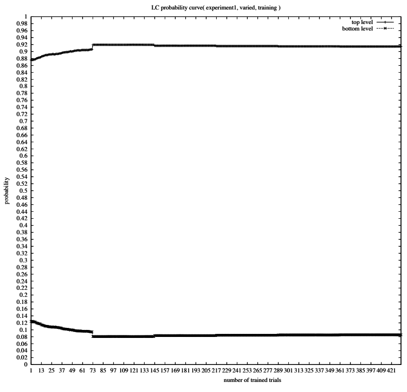


The learning curve during the simulation of Rabinowitz and Goldberg (1995).

Simulation: combination probabilities



The combination probability curve of the consistent group during training in the simulation.



The combination probability curve of the varied group during training in the simulation.

Process Control Tasks: Stanley et al. (1989)

— a system to be controlled: $P = 2 * W - P_1 + N$
— two versions (sugar vs. person)
— 12 levels of input and output
— 4 groups of subjects: original, memory training, simple rule, control

Human Data

	Sugar Task	Person Task
control	1.97	2.85
original	2.57	3.75
memory training	4.63	5.33
simple rule	4.00	5.91

— ANOVA

Simulation of Stanley et al. (1989) (with 2-step window and QBP)

Human Data

	Sugar Task	Person Task
control	1.97	2.85
original	2.57	3.75
memory training	4.63	5.33
simple rule	4.00	5.91

Model Data

	Sugar Task	Person Task
control	1.92	2.62
original	2.77	4.01
memory training	4.45	5.45
simple rule	4.80	5.65

Mean-Squared Errors

	Total	Sugar Task	Person Task
IDN+RER+IRL	0.113	0.178	0.048

Interpretation: interaction of the two levels

Finer-grained analysis:

Model Data (IDN+RER)

	Sugar Task	Person Task
control	1.55	1.89
original	1.60	1.95
memory training	3.77	4.15
simple rule	4.08	4.45

Model Data (IDN+IRL)

	Sugar Task	Person Task
control	2.10	2.65
original	3.45	4.68
memory training	4.71	5.80
simple rule	5.06	6.29

Mean-Squared Errors

	Total	Sugar Task	Person Task
IDN+RER	1.231	0.466	1.996
IDN+IRL	0.384	0.485	0.283

Model Data (IDN+RER)

	Sugar Task	Person Task
control	1.68	1.81
original	1.64	1.96
memory training	4.23	4.46
simple rule	4.72	4.87

Model Data (IDN+IRL)

	Sugar Task	Person Task
control	2.23	2.76
original	3.43	4.55
memory training	4.55	5.63
simple rule	4.86	5.63

Mean-Squared Errors

	Total	Sugar Task	Person Task
IDN+RER	1.016	0.407	1.624
IDN+IRL	0.285	0.387	0.184

— Many other probes

Accounting for Cognitive Data

— In all of these cases, simulation based on CLARION forced one to think in terms of process details

E.g., in the simulation of process control tasks, we investigated detailed computational processes involved in performing this task, in particular the two different explicit learning processes, and generated some conjectures regarding their relative importance.

— The use of the CLARION cognitive architecture provides a deeper level of scientific explanations

E.g., in simulating the alphabetic arithmetic task, explanations were provided in terms of action-centered knowledge or non-action-centered knowledge, in terms of explicit knowledge or implicit knowledge, or in terms of activations of representational units, and so on. They were deeper because the explanations were centered on lower-level mechanisms and processes

— Because of the nature of deeper explanations, this style of theorizing is also

more likely to lead to unified explanations for a large variety of data and/or phenomena

For example, all the afore-mentioned tasks have explained computationally in a unified way in CLARION

A Model for Autonomous Intelligent Systems

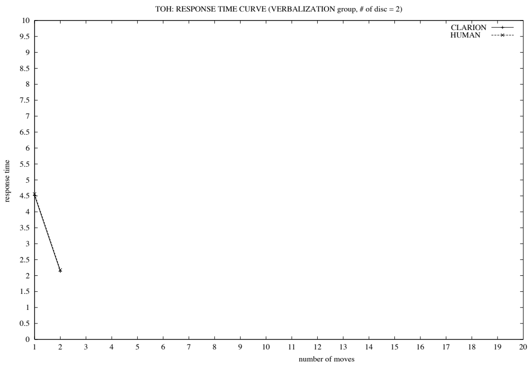
— CLARION: a model for building (autonomous) intelligent systems?

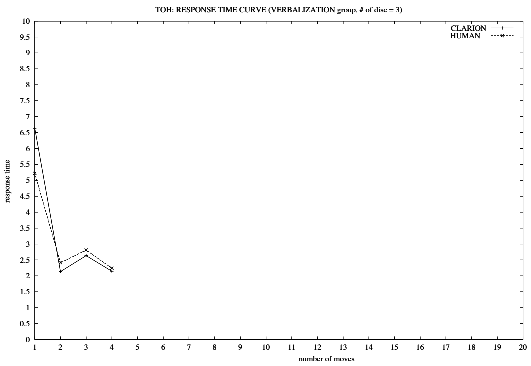
— We applied CLARION to a few reasonably interesting tasks

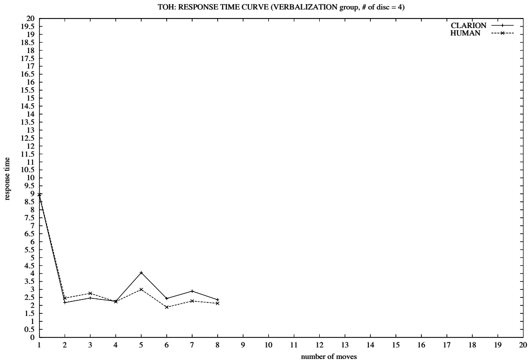
— Tasks: learning to play Tower of Hanoi and learning minefield navigation

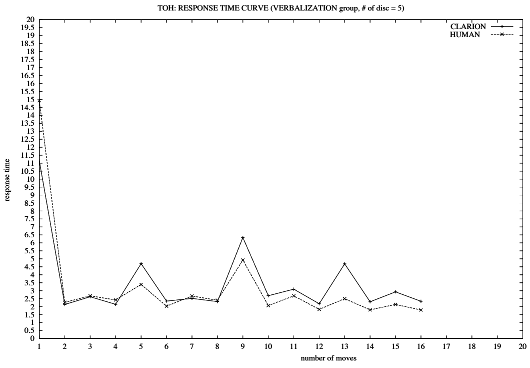
— Several different learning settings were used: bottom-up learning, top-down learning, and their combinations

Tower of Hanoi: Sun and Zhang (2004, *Cognitive Systems Research*)



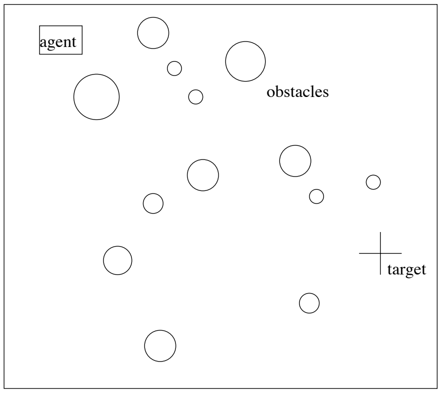


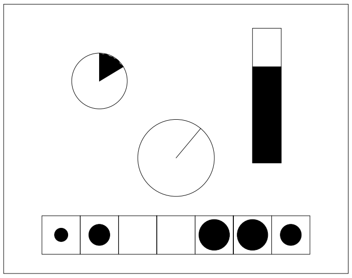



TOH: RESPONSE TIME CURVE (VERBALIZATION group, # of disc = 5)
CLARION
HUMAN
response time
20
19.5
19
18.5
18
17.5
17
16.5
16
15.5
15
14.5
14
13.5
13
12.5
12
11.5
11
10.5
10
9.5
9
8.5
8
7.5
7
6.5
6
5.5
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
number of moves

Minefield Navigation

— Limited information through instruments

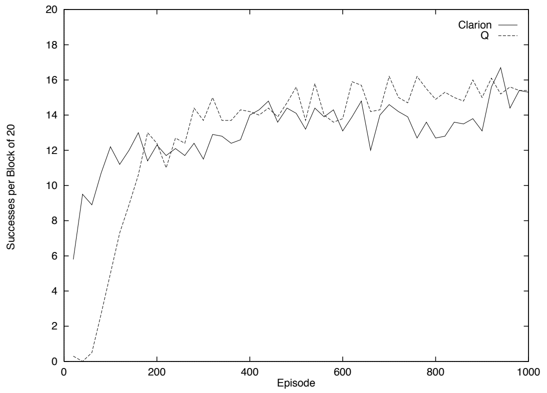




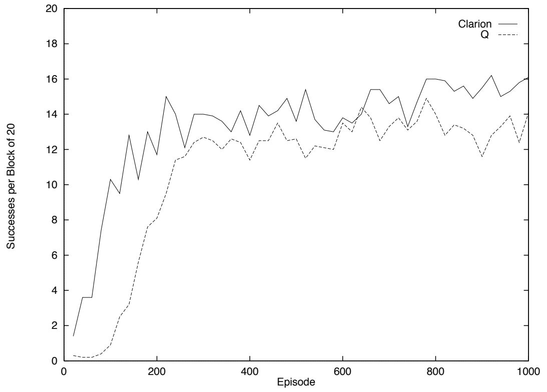
— Severe time pressure; no time for reasoning, episodic memory retrieval, and other slow processes

— Starting from scratch, without any a priori knowledge

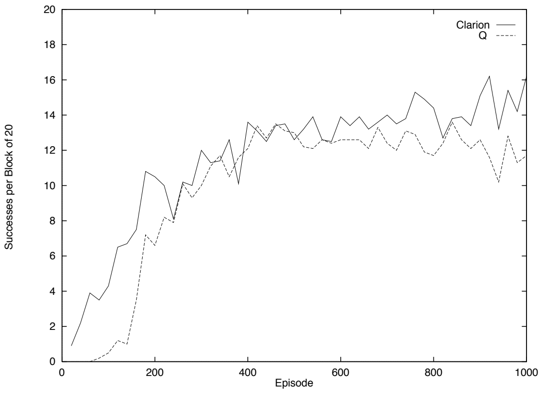
Minefield Navigation: Comparisons with regular reinforcement learning; Sun and Peterson (1998; *IEEE TNN*)



The 10-mine learning curves



The 30-mine learning curves



The 60-mine learning curves

The Challenges Ahead

In general, building cognitive architectures is an extremely difficult task:

(1) A cognitive architecture needs to be compact but yet comprehensive in scope

(2) It needs to remain simple yet capture a wide range of empirical data accurately

(3) It needs to be computationally feasible but also consistent with psychological theories

(4) It needs somehow to sort out and incorporate the myriad of incompatible psychological theories in existence, and so on

The Challenges Ahead

— Challenges from cognitive science

— Challenges to/from AI/CI

— Challenges from social simulation

The Challenges from Cognitive Science

— To integrate a broad range of cognitive functionalities, thus going against the trend of increasing specialization

— To fit pieces together smoothly

— Developing integrative cognitive architectures is thus a major challenge and a major opportunity in cognitive science

The Challenges from Cognitive Science

— In developing cognitive architectures, we need to come up with and follow a broad set of desiderata

— For example, in Anderson and Lebiere (2001) a set of desiderata proposed by Newell (1990) was used to evaluate ACT-R versus connectionist models.

These desiderata include flexible behavior, real-time performance, adaptive behavior, vast knowledge base, dynamic behavior, knowledge integration, natural language, learning, development, evolution, and brain realization

— In Sun (2004, *Philosophical Psychology*), a broader set of desiderata was proposed and used to evaluate a wider set of cognitive architectures

These desiderata include ecological realism, bio-evolutionary realism, cognitive realism, and many others

— The advantages of coming up with and applying a broad set of desiderata include (1) avoiding overly narrow models, (2) avoiding missing certain crucial functionalities, and (3) avoiding potentially inappropriate approaches or techniques in implementing cognitive architectures

— Related to that, some general architectural principles need also be examined

— It is a challenge to methodically explore such issues and reach reasonable conclusions

The Challenges from Cognitive Science

— Complex models have always invoked suspicion in psychology

Miller et al (1960): “A good scientist can draw an elephant with three parameters, and with four he can tie a knot in its tail. There must be hundred of parameters floating around in this kind of theory and nobody will ever be able to untangle them”.

— Counter-arguments can be advanced on the basis of the necessity of having complex models in understanding the mind

Miller et al (1960), Newell (1990), Sun (2002), and so on.

— However, over-generality, beyond what is minimally necessary, is always a danger

Models may account for a large set of data because of their extreme generality, rather than capturing any deep structures and regularities underlying cognitive processes

— This situation is to be avoided, by adopting a broad perspective, and by adopting a multi-level framework to fully exploit all available information and constraints

Sun et al (2005, *Philosophical Psychology*)

The Challenges from Cognitive Science

— The validation of process details of a cognitive architecture against empirical (psychological) data

There have been too many instances in the past that research communities rushed into some particular model or some particular approach toward modeling cognition and human intelligence, without knowing exactly how much of the approach or the model was veridical

— Painstakingly detailed work needs to be carried out before sweeping claims can be made

— Validation of cognitive architectures poses a serious challenge, because of a myriad of mechanisms involved in cognitive architectures, and their variety and complexity

The Challenges from/to Computational Intelligence

Langley and Laird (2003):

(1) generality, versatility, and taskability

(2) both optimality and scalability (time/space complexity)

(3) both reactivity and goal-directed behavior

(4) both autonomy and cooperation

(5) adaptation, learning, and behavioral improvements

and so on

The Challenges from/to Computational Intelligence

— To develop better, more realistic cognitive architectures, we need better algorithms

for various functionalities such as information filtering, encoding, learning, information retrieval, reasoning, decision making, problem solving, communication, and so on.

— Only on the basis of such key algorithms that are continuously improving, we may build better cognitive architectures correspondingly

In particular, we need better natural language processing capabilities, more efficient planning algorithms, more powerful learning algorithms, and so on.

— These are significant challenges from the field of cognitive architectures to AI/CI researchers

The Challenges from/to Computational Intelligence

— Various pieces have been, or are being, developed by various subfields of AI/CI

— AI/CI researchers also need to develop better computational methods (algorithms) for putting the pieces together to form a better architecture

— The challenge is to continuously improving upon the state of the art and to come up with architectures that better and better mirror the human mind and serve a variety of application domains at the same time

The Challenges from/to Computational Intelligence

— Cognitive architectures need to find both finer and broader applications, that is, both at lower levels and at higher levels

For example, some cognitive architectures found applications in large-scale simulation at a social, organizational level

For another example, some other cognitive architectures found applications in interpreting not only psychological data but also neuroimaging data (at a biological level)

— Pew & Mavor (1998) and Ritter et al (2003) provided some examples of (potential) applications of cognitive architectures

— Inevitably, this issue will provide impetus for future research (applied as well as theoretical) in cognitive architectures, and in cognitive modeling in general

Concluding Remarks

— Progress has been made in advancing the research on cognitive architectures

— There is still a long way to go

— An example cognitive architecture presented

— But need to explore more fully the space of possible cognitive architectures

— Also need to enhance the functionalities of cognitive architectures so that they can have the full range of intelligence and cognitive capabilities

— Many challenges and issues need to be addressed

— Profound impact in the future expected