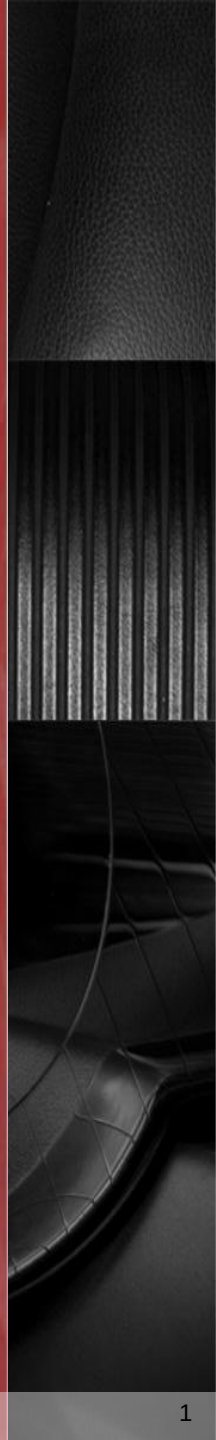
The CLARION Cognitive Architecture: A Tutorial

Ron Sun, Nick Wilson, Michael Lynch, Sébastien Hélie
Cognitive Science, Rensselaer Polytechnic Institute



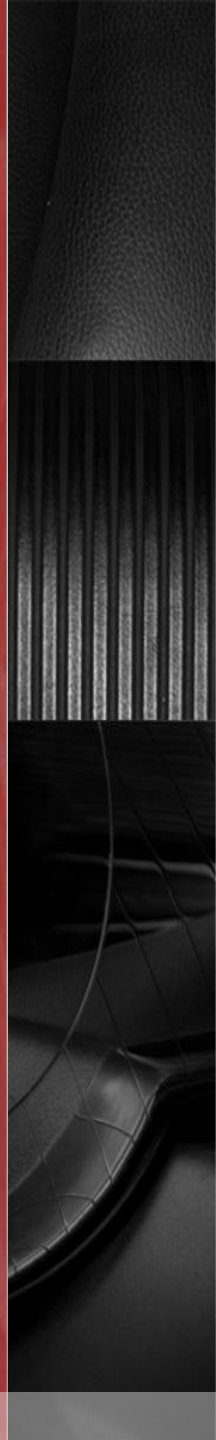
Outline of This Tutorial

- Introduction (40 min.)
  - Ron Sun, Nick Wilson, Michael Lynch, Sébastien Hélie
- The Action-Centered Subsystem (120 min.)
  - Nick Wilson, Sébastien Hélie, Ron Sun, Michael Lynch
- The Non-Action-Centered Subsystem (80 min.)
  - Sébastien Hélie, Ron Sun, Nick Wilson
- The Motivational Subsystem and Meta-Cognitive Subsystem (100 min.)
  - Nick Wilson, Ron Sun, Michael Lynch
- Conclusion (20 min.)
  - Nick Wilson, Michael Lynch, Ron Sun

The CLARION Cognitive Architecture: A Tutorial

Part 1 – Introduction

Ron Sun, Nick Wilson, Michael Lynch, Sébastien Hélie
Cognitive Science, Rensselaer Polytechnic Institute



What is a Cognitive Architecture?

A cognitive architecture is a broadly-scoped, domain-generic computational cognitive model, capturing the essential structures and processes of the mind, to be used for a broad, multiple-level, multiple-domain analysis of behavior.

See Sun (2004, Philosophical Psychology)

What is a Cognitive Architecture?

- Architecture of a building: overall design and overall framework, as well as roofs, foundations, walls, windows, floors, and so on
- Cognitive architecture: overall structures: essential divisions of modules, essential relations between modules; basic representations, essential algorithms, and a variety of other aspects within modules
- Componential processes of cognition
- Relatively invariant across time, domain, and individual
- Structurally and mechanistically well defined

What is a Cognitive Architecture?

Functions (in relation to cognitive science and in relation to AI):

- To provide an essential framework to facilitate more detailed modeling and exploration of various components of the mind -- mechanisms and processes …
  - … specifying computational models of cognitive mechanisms and processes
  - … embodying theories/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 -- more cognitively grounded intelligent systems

Why are Cognitive Architectures Important for Cognitive Science?

- Psychologically oriented cognitive architectures: “intelligent” systems that are cognitively-psychologically realistic; detailed cognitive-psychological theories that have been tested through capturing and explaining psychological data; and so on
- They help to shed new light on human cognition and therefore they are useful tools for advancing the science of cognition-psychology
- They may 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 and in terms of mechanistic detail
- Require that important elements of a theory 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 cognitive phenomena
- Developing generic models of cognition (capable of a wide range of cognitive functionalities) helps to avoid the myopia of narrowly-scoped research

see *Newell (1990)* and *Sun (2002, book published by Erlbaum)*

Still Room for Grand Theories?

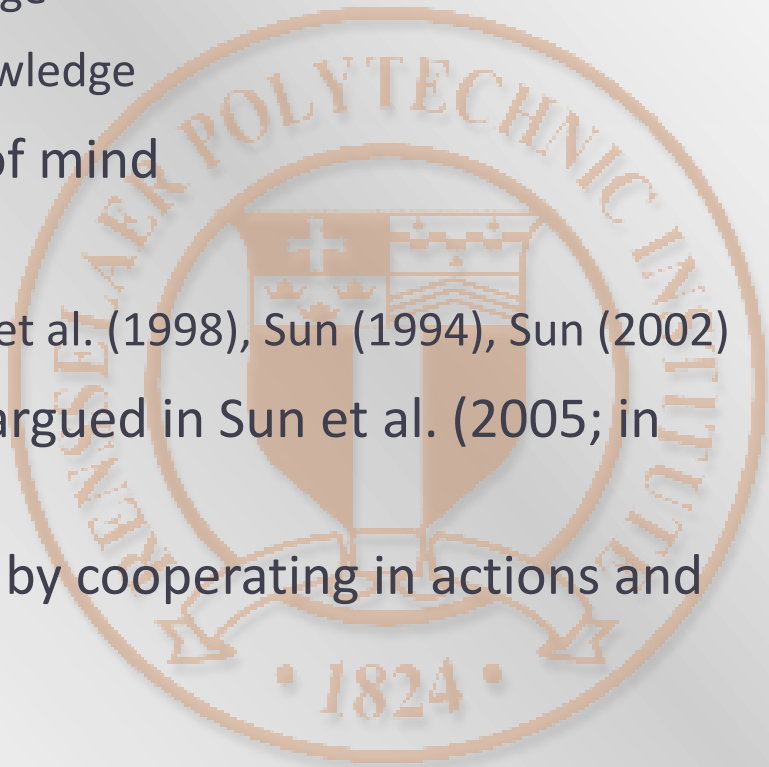
- Some have claimed that fundamental scientific discovery and grand scientific theorizing have become a thing of the past. What remains to be done is filling in details
- Researchers in cognitive science are pursuing integrative approaches that explain data in multiple levels, domains, and functionalities
- Significant advances may be made through hypothesizing and confirming deep-level principles that unify superficial explanations across multiple domains
- Cognitive architectures can be the basis of such unified theories in cognitive science (see, e.g., *Sun, 2002, the Erlbaum book*)

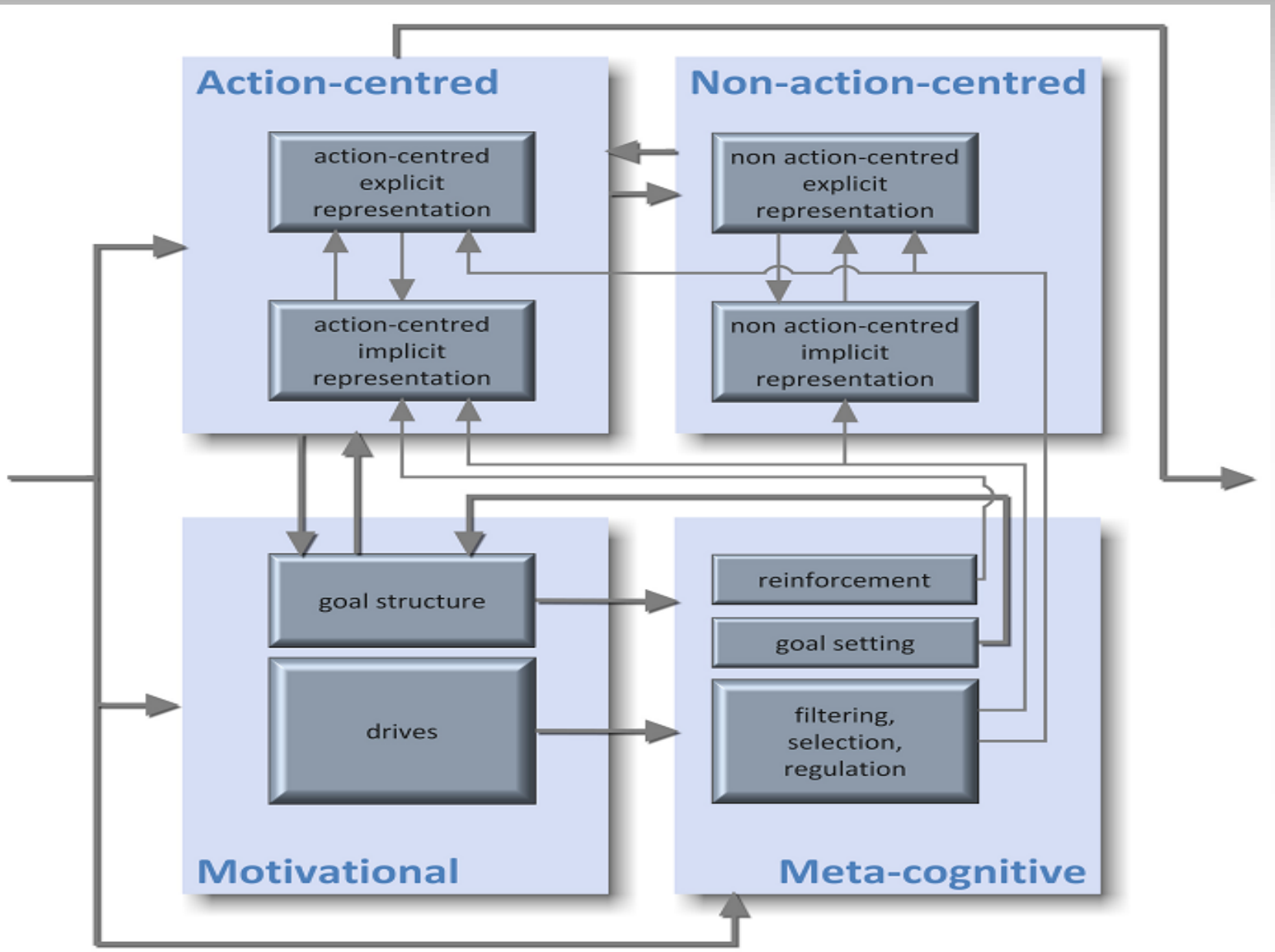
CLARION: An Example of a Cognitive Architecture

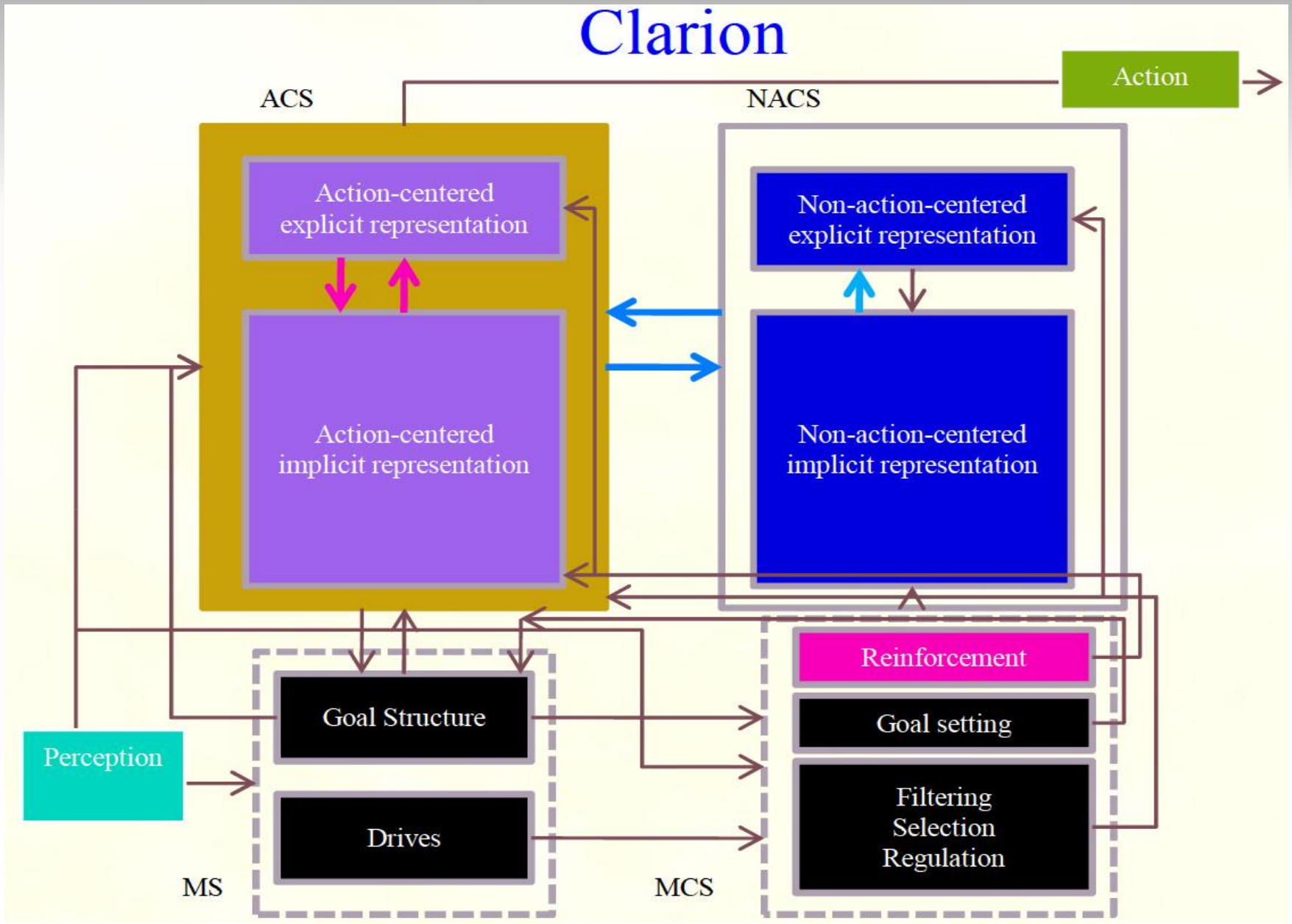
- An integrative cognitive architecture, consisting of a number of distinct but symbiotic subsystems (with critical mutual dependencies and complex interactions)
- A dual-representational structure in each subsystem (implicit versus explicit representations)
- Its subsystems include:
  - the Action-Centered Subsystem (the ACS) → procedural
  - the Non-Action-Centered Subsystem (the NACS) → declarative
  - the Motivational Subsystem (the MS), and
  - the Meta-Cognitive Subsystem (the MCS)

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
- Essentially, it is a dual-process theory of mind
  - Evans and Frankish (2009)
  - Reber (1989), Seger (1994), Cleeremans et al. (1998), Sun (1994), Sun (2002)
- Duality of representation: extensively argued in Sun et al. (2005; in Psychological Review)
- The two "levels" interact, for example, by cooperating in actions and in learning




Action-centred
Non-action-centred
action-centred explicit representation
non action-centred explicit representation
action-centred implicit representation
non action-centred implicit representation
goal structure
drives
reinforcement
goal setting
filtering, selection, regulation
Motivational
Meta-cognitive


Clarion
Action
ACS
NACS
Action-centered
explicit representation
Action-centered
implicit representation
Non-action-centered
explicit representation
Non-action-centered
implicit representation
Reinforcement
Goal Structure
Goal setting
Perception
Drives
Filtering
Selection
Regulation
MS
MCS

Essential Characteristics

- The dichotomy of implicit and explicit processes:

  justifications later

- The focus on the cognition-motivation-environment interaction:

  justifications later

- The constant interaction of multiple subsystems and modules:

  involving implicit cognition, explicit cognition, motivation, meta-cognition, and so on; complex interactions and mutual dependencies

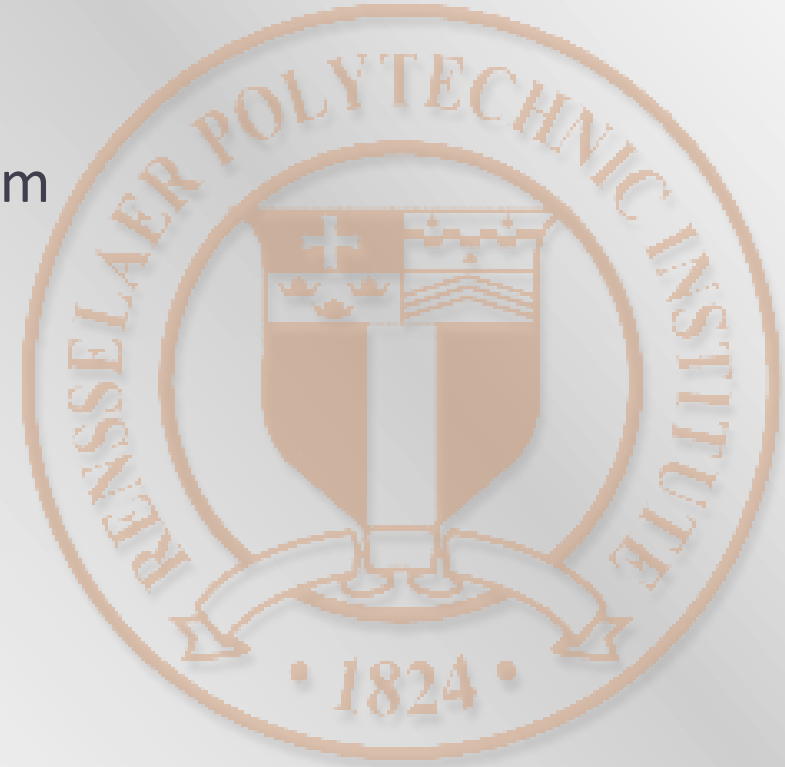
- Autonomous and bottom-up/top-down learning:

  CLARION can learn on its own, regardless of whether there is a priori or externally provided domain knowledge, while it does not exclude innate biases, innate behavioral propensities, prior knowledge, etc.

see *Sun (2004, Philosophical Psychology)*

Sketching Quickly Some Details of the Subsystems

- The Action-Centered Subsystem
- The Non-Action-Centered Subsystem
- The Motivational Subsystem
- The Meta-Cognitive Subsystem



Sketching Some Details of the Subsystems

- The Action-Centered Subsystem
- The Non-Action-Centered Subsystem
- The Motivational Subsystem
- The Meta-Cognitive Subsystem

The Action-Centered Subsystem

- In the bottom level of the action-centered subsystem, implicit reactive action routines are formed/learned:
  - Values and reinforcement learning
  - Modularity
  - Essential to and primary in cognition (Sun, 2002)
- See: *Sun et al (2001, Cognitive Science)* and *Sun (2003, Technical Specification)* for details

The Action-Centered Subsystem

- In the top level of the action-centered subsystem, explicit action knowledge is captured in the form of explicit symbolic rules and learned through a variety of means
- With regard to explicit knowledge at the top level:
  - Bottom-up learning
  - Top-down learning
  - Independent hypothesis testing learning of explicit knowledge
  - Other forms of learning

The Action-Centered Subsystem

Autonomous generation of grounded 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 interactions with the world, the agent verifies and modifies 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.

- Statistical measures

The Action-Centered Subsystem

- Bottom-up learning: A kind of “rational” (explicit) reconstruction of implicit knowledge
- After explicit rules have been learned, a variety of explicit reasoning may be performed — Sun (2003)
- Explicit knowledge at the top level: Enhance skilled performance, facilitate verbal communication, and so on
- Learning explicit representations at the top level can be useful in enhancing learning at the bottom level — *Sun et al. (2001); Sun et al. (2005)*

The Action-Centered Subsystem

Assimilation of externally given conceptual structures

- CLARION can learn even when no a priori or externally provided explicit 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

Sketching Some Details of the Subsystems

- The Action-Centered Subsystem
- **The Non-Action-Centered Subsystem**
- The Motivational Subsystem
- The Meta-Cognitive Subsystem

The Non-Action-Centered Subsystem

- Representing general knowledge about the world – that is, declarative knowledge
  - the “semantic” memory (Quillian, 1968)
  - the episodic memory (Tulving, 1972)
- Performing various kinds of memory retrievals and inferences
- Under the direction of the action-centered subsystem (through its actions)

The Non-Action-Centered Subsystem

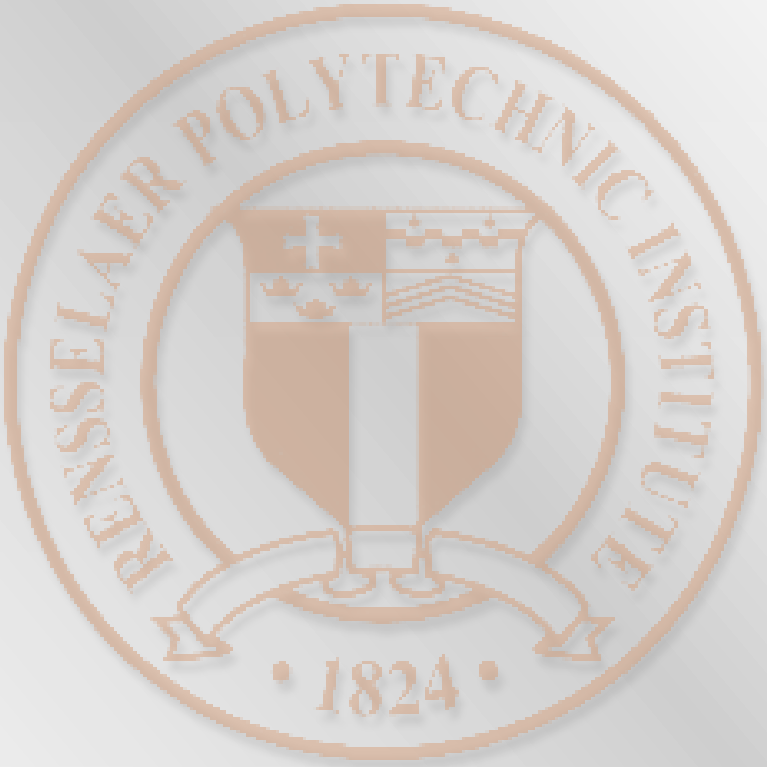
- At the bottom level: "associative memory" networks encode implicit non-action-centered knowledge, with distributed representation of (micro)features
- At the top level: explicit non-action-centered knowledge is encoded:
  - symbolic/localist representation of concepts, i.e., chunk nodes
  - A node is set up in the top level to represent a chunk (a concept), and connects to its corresponding (micro)features (distributed representation) in the bottom level
- At the top level, links between chunk nodes encode associations between pairs of chunks (concepts) — explicit 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 human everyday commonsense reasoning patterns (including "inheritance reasoning")
  - See *Sun (1994, book published by Wiley)*, and *Sun (1995, Artificial Intelligence)*

The Non-Action-Centered Subsystem

- Bottom-up learning
- Top-down learning
- Other forms of learning



Sketching Some Details of the Subsystems

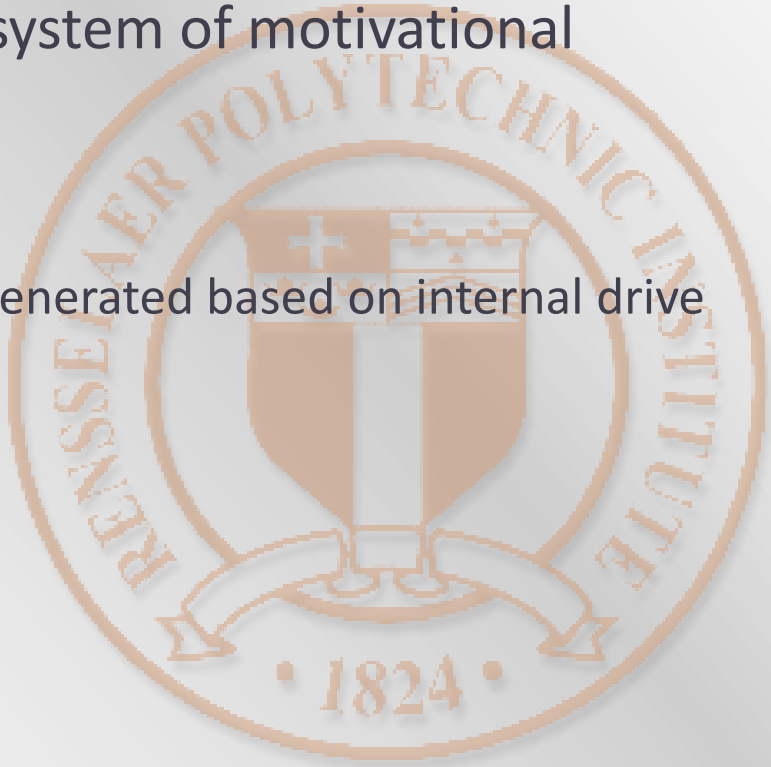
- The Action-Centered Subsystem
- The Non-Action-Centered Subsystem
- **The Motivational Subsystem**
- The Meta-Cognitive Subsystem

The Motivational Subsystem

- Concerned with why an agent does what it does.
- Simply saying that an agent chooses actions to maximize gains, rewards, reinforcements, or payoffs leaves open the question of what determines these things
- Drives and their interactions lead to goals and actions (Murray, 1938; Toates, 1986)

The Motivational Subsystem

- 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 representations:
  - Explicit goals vs. drive activations
  - The explicit goals of an agent may be generated based on internal drive activations



The Motivational Subsystem

- Low-level primary drives (mostly physiological): hunger, thirst, physical danger, ...
- High-level primary drives (mostly social): seeking of social approval, striving for social status, desire for reciprocation, ...
- Secondary (derived) drives
  - 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, e.g., through externally given instructions

Sketching Some Details of the Subsystems

- The Action-Centered Subsystem
- The Non-Action-Centered Subsystem
- The Motivational Subsystem
- **The Meta-Cognitive Subsystem**

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)

  Studied extensively in experimental psychology

  See, e.g., Schwartz and Shapiro (1986), Metcalfe and Shimamura (1994), Reder (1996), Mazzoni and Nelson (1998)

- Regulates not only goal structures but also cognitive processes *per se*.

The Meta-Cognitive Subsystem

(1) behavioral aiming:

- setting of reinforcement functions
- setting of goals

(2) information filtering:

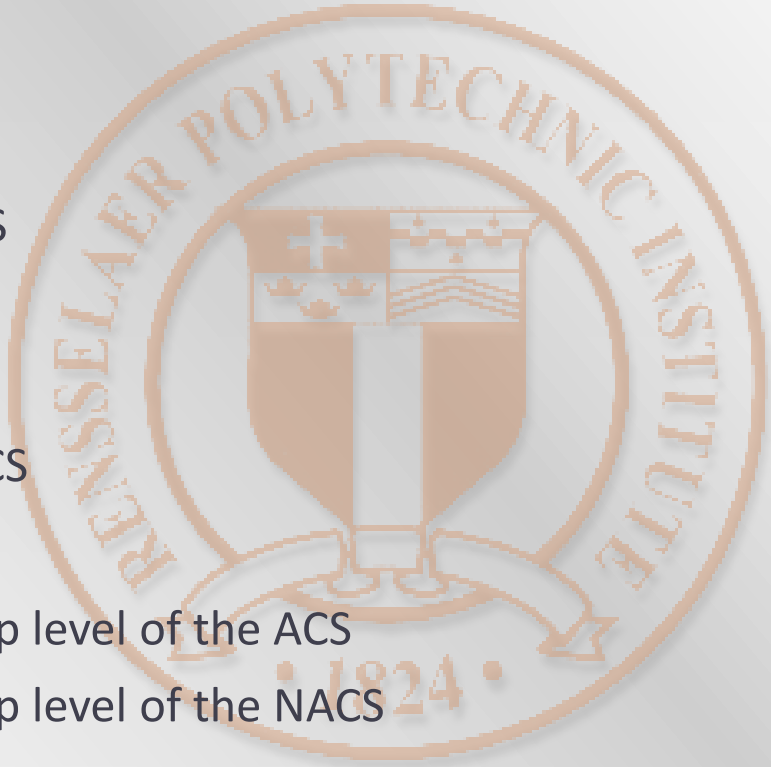
- focusing of input dimensions in the ACS
- focusing of input dimensions in the NACS

(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



The Meta-Cognitive Subsystem

(5) outcome selection:

- selection of output dimensions in the ACS
- selection of 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

(7) parameters of the ACS and the NACS:

- setting of parameters for the bottom level of the ACS
- setting of parameters for the bottom level of the NACS
- setting of parameters for the top level of the ACS
- setting of parameters for the top level of the NACS

Too Many Mechanisms?

- Are there too many specialized mechanisms?
  - General "semantic" memory, in both implicit and explicit forms (in the non-action-centered subsystem, for general knowledge)
  - Episodic memory (in the non-action-centered subsystem)
  - Procedural memory, in both implicit and explicit forms (in the action-centered subsystem)
  - Working memory (in the action-centered subsystem)
  - Goal structures (in the action-centered subsystem; a part of WM)
- In general, CLARION is grounded in existing psychological theories, constitutes a comprehensive psychological theory by itself, is reasonably compact, and matches a wide range of psychological data (Sun, 2002, 2003)

Differences with ACT-R

- CLARION makes a principled distinction between explicit and implicit processes/knowledge/learning:
  - ACT-R does not directly capture the distinction and the interaction between implicit and explicit processes;
  - ACT-R provides no direct explanation of synergy effects between the two types of processes/knowledge/learning (Sun et al., 2005).
- ACT-R is not meant for autonomous learning, without a lot of a priori knowledge; it does not directly capture the psychological process of bottom-up learning as CLARION does.
- CLARION is capable of automatic and 'effortless' similarity-based reasoning, while ACT-R has to use cumbersome pair-wise similarity relations.
- CLARION has a general functional approximation capability (in its bottom level), while ACT-R does not.

Differences with ACT-R

- In ACT-R, there is no built-in modeling of motivational processes (as in CLARION) – goals are externally set and directly hand-coded.
- In ACT-R, there is no built-in sophisticated meta-cognitive process (as in CLARION). (But some recent attempts.)
- ACT-R has some detailed sensory-motor modules that CLARION currently does not include.
- CLARION and ACT-R often account for different tasks, although there have been some overlaps also.

Differences with SOAR

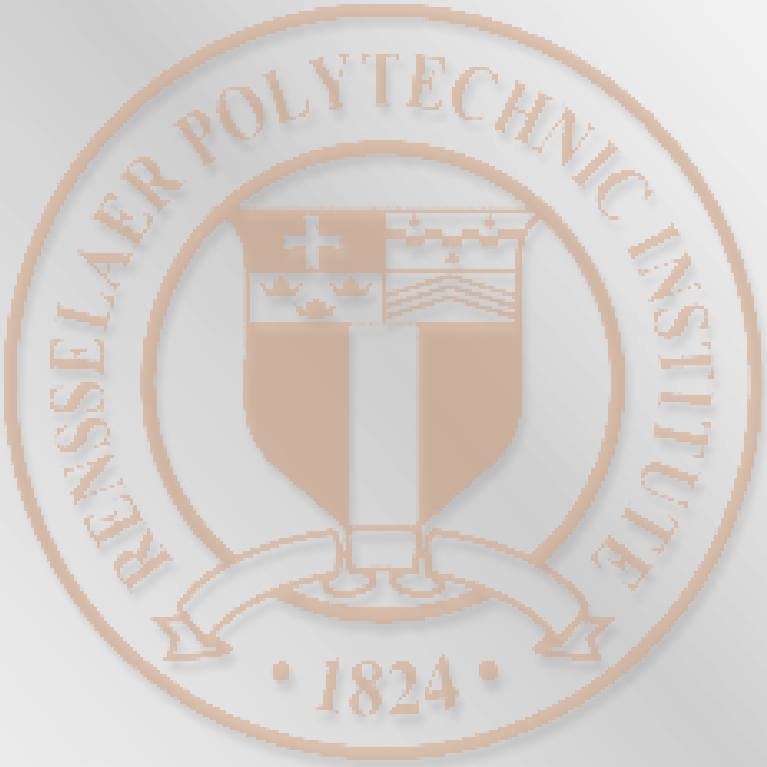
- In Soar, a large amount of initial (a priori) knowledge is required, and thus no autonomous learning and no bottom-up learning.
- Soar makes no distinction between explicit and implicit knowledge and learning (and its learning is based on specialization using only symbolic representations).
- In Soar, there is no built-in modeling of the psychological process of the interaction and synergy between explicit and implicit processes.
- In Soar, there is no distinction between symbolic/localist and distributed representations. Nor is there general function approximation capability.
- It does not embody similarity-based reasoning processes directly.
- In Soar, there is no built-in motivational process. Nor is there built-in sophisticated meta-cognitive process.

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)

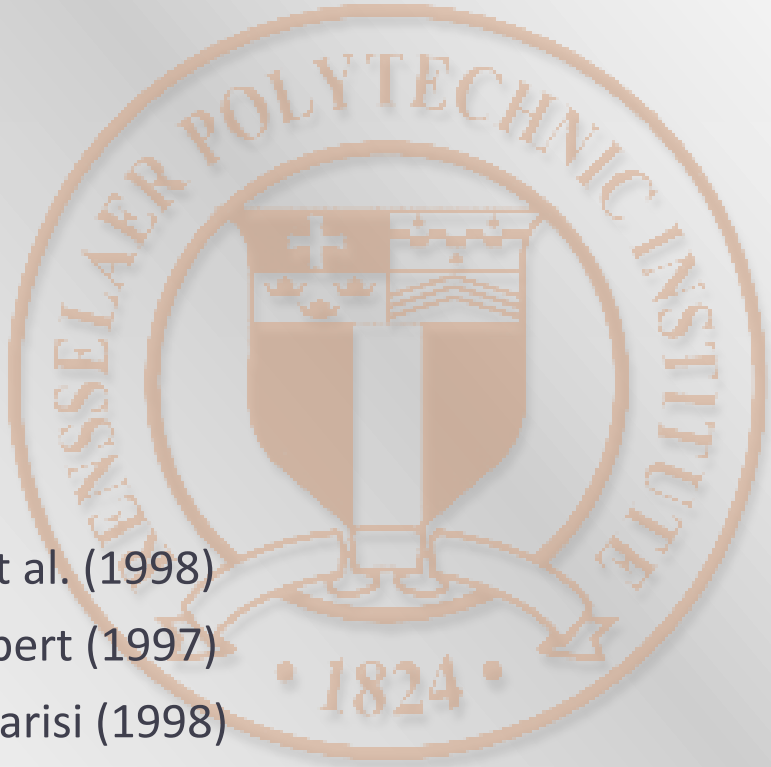
Accounting for Cognitive Data: Past simulations using CLARION

- Minefield navigation
  - Sun et al. (2001)
- Tower of Hanoi
  - Gagne and Smith (1962)
- Categorical inference tasks
  - Sloman (1998)
- Discovery tasks
  - Bowers et al. (1990)



Accounting for Cognitive Data: Past simulations using CLARION

- "Lack of knowledge" inferences
  - Gentner and Collins (1991)
- Meta-cognitive monitoring
  - Metcalfe (1986)
- Motivational processes
  - Lambert et al. (2003)
  - Beilock et al. (2004)
  - Beilock and Carr (2001)
- Social simulations
  - Organizational decision making: Carley et al. (1998)
  - Scientific productivity: Simon (1957); Gilbert (1997)
  - Survival of tribal societies: Cecconi and Parisi (1998)



Accounting for Cognitive Data: Past simulations using CLARION

- Creative problem solving
  - Smith and Vela (1991)
  - Yaniv and Meyer (1987)
  - Durso et al. (1994)
  - Schooler et al. (1993)
- Moral judgment
- Personality
- Focus: capturing the interaction, and the resulting synergy, emphasizing bottom-up learning; interaction of cognition, motivation, and meta-cognition.

Psychological Justifications and Implications of CLARION

- R. Sun (2013). ***Anatomy of Mind***. Oxford University Press, New York.
- R. Sun (2002). ***Duality of the Mind***. Lawrence Erlbaum Associates, Mahwah, NJ.
- R. Sun (1994). ***Integrating Rules and Connectionism for Robust Commonsense Reasoning***. John Wiley and Sons, New York.

- S. Hélie and R. Sun (2010). Insight, incubation, and creative problem solving: A unified theory and a connectionist model. *Psychological Review, 117(3)*, 994-1024.
- R. Sun, P. Slusarz, and C. Terry (2005). The interaction of the explicit and the implicit in skill learning: A dual-process approach. *Psychological Review, Vol.112, No.1, pp.159-192*.
- R. Sun, E. Merrill, and T. Peterson (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science, Vol.25, No.2*, pp.203-244.
- R. Sun (1995). Robust reasoning: Integrating rule-based and similarity-based reasoning. *Artificial Intelligence. Vol.75, No.2*, pp.241-296.

Technical Details of CLARION

- ***A Detailed Specification of CLARION 5.0***. Technical report, RPI. (It contains detailed technical specifications of CLARION 5.0.)
  - Addendum 1: The enhanced description of the motivational subsystem.
  - Addendum 2: The enhanced description of similarity-based reasoning.
  - Addendum 3: The properties of the CLARION-H implementation.
  - Addendum 4: Q and A.
- will be updated and published by Oxford University Press, 2014
- A much simplified description of CLARION 5.0, written by a student as a project report (which only provides some general ideas): A Simplified Introduction to CLARION 5.0. Technical report. 2004.

Conclusion: What is CLARION?

- A comprehensive theory of the mind (i.e., cognition as broadly defined)
- A conceptual framework for analyzing cognition/psychological processes (various functionalities and tasks)
- A computational modeling framework for simulating cognitive-psychological data
- A set of simulation programming tools (Java packages [5.0, 6.0]; C# packages [6.1])

End of Part 1: Introduction

- Any general questions at this point?
- Note: details to follow