Cognitive Architectures

Knowledge, reasoning, and decision making

General presentation

Othalia Larue and Pierre Poirier
ISC Summer School 2026

Unified theories of cognition as cognitive architectures
A problem and a solution

Playing 20 questions with nature (Newell 1973)

- Nature versus nurture
- Peripheral versus central processing
- Innate versus learned rules / representation / architecture
- Conscious versus unconscious
- Sequential versus parallel processing
- Pre-attentive versus attentive
- Contextual versus independent interpretation
- Top-down versus bottom-up processing

Current (i.e., 1973) efforts are not contributing to an integrated approach to cognition, leading to theoretical myopia

- Symbolic approach
- Connectionist approach
- Dynamical approach
- Modular, massively modular, non-modular approaches
- Dual-process approach
- Internalist approach
- Externalist approach
- Embodied approach
- Enactive approach
- Ecological approach
- Predictive approach

Unified theories of cognition as cognitive architectures
A problem and a solution

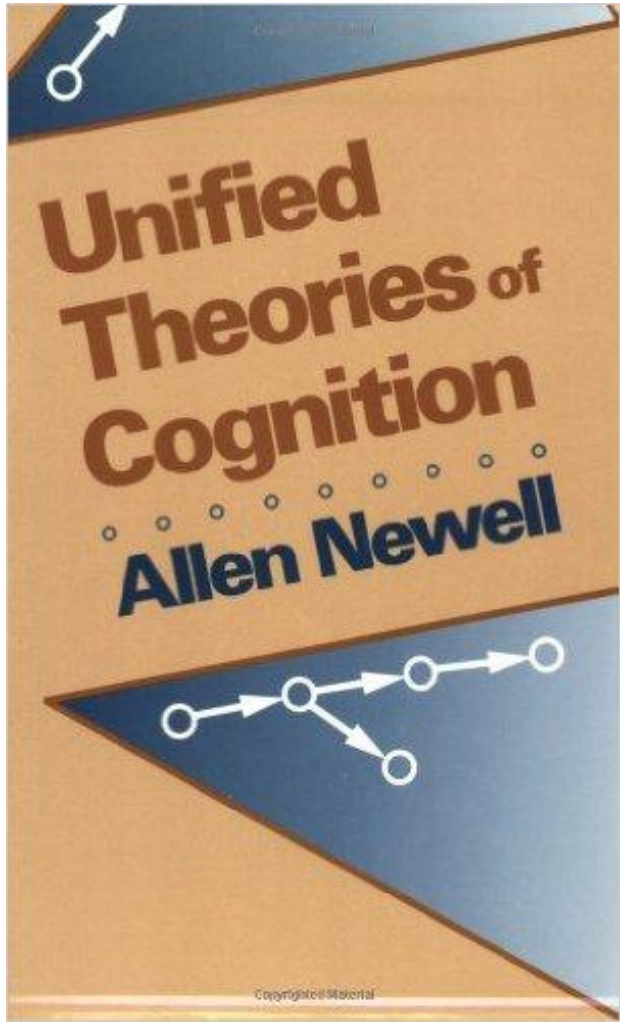
One Program for Many Tasks

The third alternative paradigm I have in mind is to stay with the diverse collection of small experimental tasks, as now, but to construct a single system to perform them all. This single system (this model of the human information processor) would have to take the instructions for each, as well as carry out the task. For it must truly be a single system in order to provide the integration that we seek.

The companion piece on productions systems (Newell, this volume, Chapter 10) in conjunction with Klahr's production system (Klahr, this volume, Chapter 11) indicates how such an endeavor might go. It is only a beginning, but it shows already a certain promise, it seems to me.

An alternative mold for such a task is to construct a single program that would take a standard intelligence test, say the WAIS or the Stanford-Binet. This is actually an enterprise that was called for much earlier (Green, 1964), but only recently has anything really stirred (Hunt, Frost, & Lunneborg, 1972).

Unified theories of cognition as cognitive architectures
A problem and a solution



BEHAVIORAL AND BRAIN SCIENCES (1992) 15, 425–492
Printed in the United States of America

Précis of *Unified theories of cognition*

Allen Newell
School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
Electronic mail: *newell@ cs.cmu.edu*

Abstract: The book presents the case that cognitive science should turn its attention to developing theories of human cognition that cover the full range of human perceptual, cognitive, and action phenomena. Cognitive science has now produced a massive number of high-quality regularities with many microtheories that reveal important mechanisms. The need for integration is pressing and will continue to increase. Equally important, cognitive science now has the theoretical concepts and tools to support serious attempts at unified theories. The argument is made entirely by presenting an exemplar unified theory of cognition both to show what a real unified theory would be like and to provide convincing evidence that such theories are feasible. The exemplar is SOAR, a cognitive architecture, which is realized as a software system. After a detailed discussion of the architecture and its properties, with its relation to the constraints on cognition in the real world and to existing ideas in cognitive science, SOAR is used as theory for a wide range of cognitive phenomena: immediate responses (stimulus-response compatibility and the Sternberg phenomena); discrete motor skills (transcription typing); memory and learning (episodic memory and the acquisition of skill through practice); problem solving (cryptarithmetic puzzles and syllogistic reasoning); language (sentence verification and taking instructions); and development (transitions in the balance beam task). The treatments vary in depth and adequacy, but they clearly reveal a single, highly specific, operational theory that works over the entire range of human cognition. SOAR is presented as an exemplar unified theory, not as the sole candidate. Cognitive science is not ready yet for a single theory – there must be multiple attempts. But cognitive science must begin to work toward such unified theories.

Keywords: artificial intelligence; chunking; cognition; cognitive science; computation; problem solving; production systems; SOAR; symbol systems

The book begins by urging on psychology unified theories of cognition:

> Psychology has arrived at the possibility of unified theories of cognition – theories that gain their power by positing a single system of mechanisms that operate together to produce the full range of human cognition.
>
> I do not say they are here, but they are within reach and we should strive to attain them.

My goal is to convince the reader that unified theories of cognition are really worth striving for – now, as we move into the nineties. This cannot be done just by talking about it. An exemplar candidate is put forth to illustrate concretely what a unified theory of cognition means and why it should be a goal for cognitive science. The candidate is a theory (and system) called SOAR (Laird et al. 1987).

The book is the written version of the *William James Lectures*, delivered at Harvard University in spring 1987. Its stance is personal, reflecting the author's thirty years of research in cognitive science, although this précis will be unable to convey much of this flavor.

Chapter 1: Introduction

The first chapter describes the enterprise. It grounds the concerns for how cognitive science should proceed by reflecting on a well-known earlier paper entitled "You can't play 20 questions with nature and win" (Newell 1973a), which even then fretted about the gap between the empirical and theoretical progress in cognitive psychology and called for more integrative theories. This book may be seen as a step toward answering that call.

The nature of theories. Chapter 1 discusses the notion of theory, to ground communication, building on some concrete examples: Fitts's Law, the power law of practice, and a theory of search in problem spaces. There is nothing special about a theory just because it deals with the human mind. It is important, however, that the theory make predictions, not the theorist. Theories are always approximate, often deliberately so, in order to deliver useful answers. Theories cumulate, being refined and reformulated, corrected and expanded. This view is Lakatosian, rather than Popperian: A science has investments in its theories and it is better to correct one than to discard it.

What are unified theories of cognition? Unified theories of cognition are single sets of mechanisms that cover all of cognition – problem solving, decision making, routine action, memory, learning, skill, perception, motor activity, language, motivation, emotion, imagining, dreaming, daydreaming, and so on. Cognition must be taken broadly to include perception and motor activity. No unified theory of cognition will deal with the full list above

© 1992 Cambridge University Press 0140-525X/92 $5.00+.00

Why strive for unified theories, beyond the apple-pie desire of all sciences to be unified? The biggest reason is that a single system (the mind) produces behavior. There are other reasons, however. Cognitive theory is radically underdetermined by data, hence as many constraints as possible are needed and unification makes this possible. A unified theory is a vehicle of cumulation simply as a theoretically motivated repository. A unified theory increases identifiability and allows theoretical constructs to be amortized over a wide base of phenomena.

Architectures. Unified theories of cognition will be formulated as architectures. The architecture of the mind is a major source of commonality of behavior, both within an individual and between individuals. The architecture is

What are cognitive architectures?
Definitions, definitions

- A cognitive architecture is a general proposal about the representations and processes that produce intelligent thought. Cognitive architectures have primarily been used to explain important aspects of human thinking such as problem solving, memory, and learning. But they can also be used as blueprints for designing computers and robots that possess some of the cognitive abilities of humans. (p.50)
  - Thagard, P. (2012). Cognitive architectures. In K. Frankish & W. Ramsay (Eds.), *The Cambridge Handbook of Cognitive Science*. Cambridge: Cambridge University Press, pp. 50-70.

- A cognitive architecture is the overall, essential structure and process of a (broadly scoped) domain generic cognitive model, used for a broad, multiple-level analysis of cognition and behavior. (p.341)
  - Sun, R. (2004). Desiderata for Cognitive Architectures. *Philosophical Psychology*, 17(3), 341-373.

What are cognitive architectures?
An important distinction

- The theoretical or conceptual architecture
  - A theory (or general conception) of the organization of cognition;
    - Ex.: Cognition is fundamentally predictive.
- The computational architecture
  - A formal model implementing that theory or conception of cognition;
    - Ex.: Cognition is a hierarchical system that performs approximate Bayesian inference through variational optimization, implemented via the exchange of predictions and prediction errors between levels of the hierarchy, with the goal of minimizing prediction error.
- The software implementation
  - A particular program realizing the computational model.
    - Ex.: https://github.com/infer-actively/pymdp (Python Markov Decision Process)

From theoretical to computational cognitive architectures
A Marr-like characterization

Level	Predictive processing
Computational goal	Minimize prediction error (or free energy)
Computational theory	Bayesian inference about hidden causes
Algorithm and representations	Variational inference and probability densities
Mechanism	Hierarchical message passing

Cognitive architectures at 40

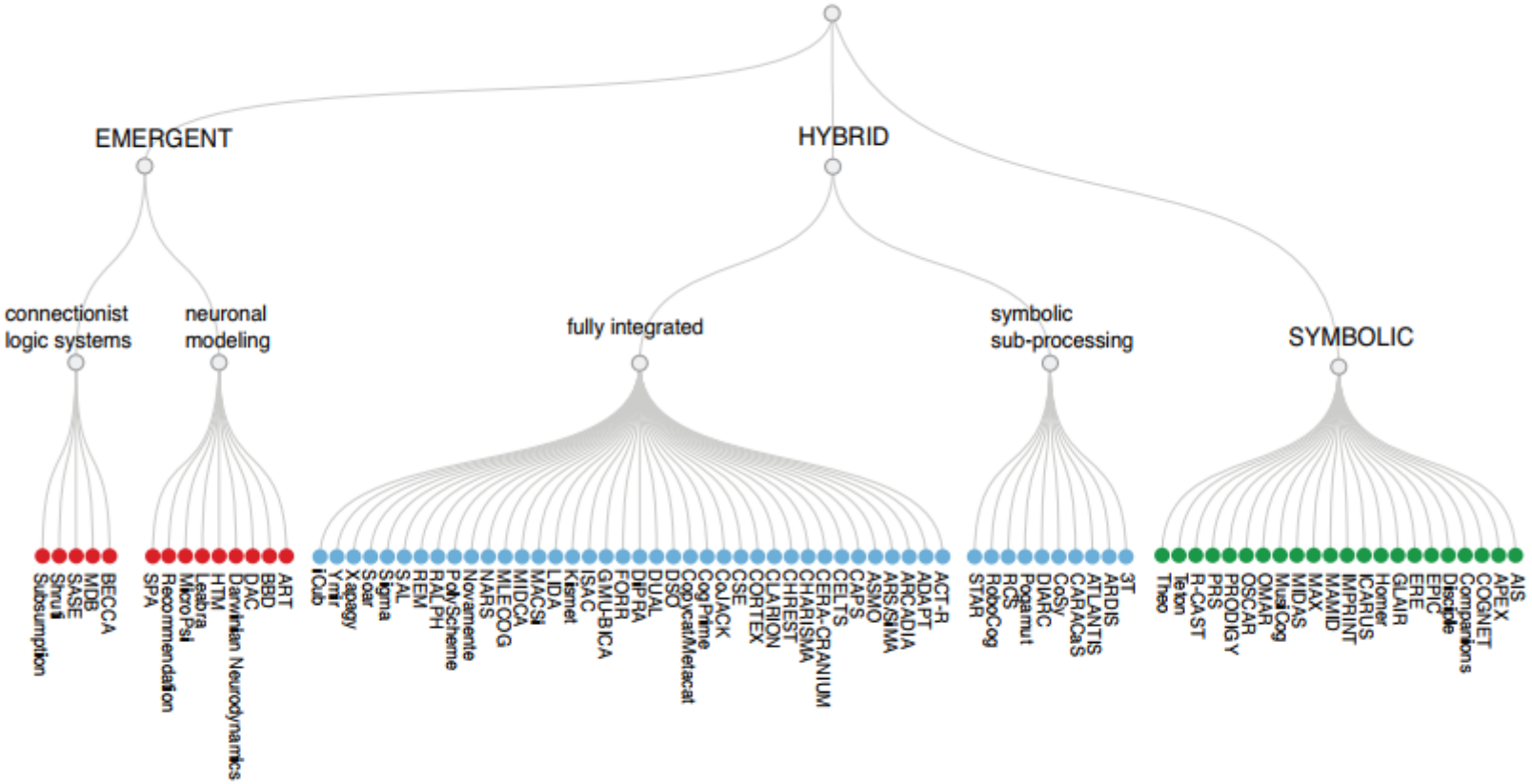


Fig. 3 A taxonomy of cognitive architectures based on the representation and processing. The order of the architectures within each group is alphabetical and does not correspond to the proportion of symbolic vs sub-symbolic elements (i.e. the spatial proximity of ACT-R and iCub to nodes representing symbolic and emergent architectures respectively does not imply that ACT-R is closer to symbolic paradigm and iCub is conceptually related to emergent architectures).

Criteria for cognitive architectures

The Newell Test for a theory of cognition (Anderson and Lebiere, 2003)

1. Behave as an (almost) arbitrary function of the environment
   –Is it computationally universal with failure?
   Classical Connectionism: Mixed; ACT-R: Better
2. Operate in real time
   –Given its timing assumptions, can it respond as fast as humans?
   Classical Connectionism: Worse; ACT-R: Best
3. Exhibit rational, that is, effective adaptive behavior
   –Does the system yield functional behavior in the real world?
   Classical Connectionism: Better; ACT-R: Better
4. Use vast amounts of knowledge about the environment
   –How does the size of the knowledge base affect performance?
   Classical Connectionism: Worse; ACT-R: Mixed
5. Behave robustly in the face of error, the unexpected, and the unknown
   –Can it produce cognitive agents that successfully inhabit dynamic environments?
   Classical Connectionism: Mixed; ACT-R: Better
6. Integrate diverse knowledge
   –Is it capable of common examples of intellectual combination?
   Classical Connectionism: Worse; ACT-R: Mixed
7. Use (natural) language
   –Is it ready to take a test of language proficiency?
   Classical Connectionism: Better; ACT-R: Worse
8. Exhibit self-awareness and a sense of self
   –Can it produce functional accounts of phenomena that reflect consciousness?
   Classical Connectionism: Worse; ACT-R: Worse
9. Learn from its environment
   –Can it produce the variety of human learning?
   Classical Connectionism: Better; ACT-R: Better
10. Acquire capabilities through development
    –Can it account for developmental phenomena?
    Classical Connectionism: Better; ACT-R: Worse
11. Arise through evolution
    –Does the theory relate to evolutionary and comparative considerations?
    Classical Connectionism: Worst; ACT-R: Worst
12. Be realizable within the brain
    –Do the components of the theory exhaustively map onto brain processes?
    Classical Connectionism: Best; ACT-R: Worse

Criteria for cognitive architectures
Other lists of criteria

- Desiderata for cognitive architectures
  - Sun, R. (2004). "Desiderata for cognitive architectures." *Philosophical Psychology* 17(3): 341-373.
- *Core Cognitive Criteria*
  - Eliasmith, C. (2013). *How to build a brain: A neural architecture for biological cognition*, Oxford University Press.

Today’s talks

1. Cognitive models of information effects (Othalia Larue)

The modern information environment has made it increasingly difficult to distinguish fact from fiction. Additionally, some information is deliberately crafted and disseminated to exploit human cognitive processes. This environment is highly complex, which has made it difficult to study using traditional experimental methods. In other areas of cognitive science, researchers have addressed similar challenges by leveraging computational cognitive models alongside experimental approaches, leading to significant advances. However, existing models in this domain have primarily focused on identifying and characterizing features of the information environment rather than explaining their effects on cognition and behavior. We shift the focus toward explanation, developing and examining different models of information-related effects.

2. Reasoning in Cognitive Architectures: From Rule-Based to Data-Driven (Iuliia Kotseruba)

Reasoning is involved in most cognitive functions, from perception to action. Although reasoning is often associated with precise logic, everyday reasoning is often illogical, irrational, and suboptimal. Reasoning from potentially incomplete and uncertain premises to this day remains an open problem in developing computational models of cognition.

In this talk I will discuss the evolution of computational solutions to the problem of human reasoning from early heuristic approaches to today’s data-driven ones. I will start by defining reasoning to distinguish it from cognition and thinking. I will then talk about theoretical reasoning (about beliefs) and practical reasoning (action selection) and aspects of both that help deal with uncertainty and requirements for physical actions. I will then examine biological and artificial mechanisms that implement these features. Lastly, I will briefly discuss how behavior modulators (e.g. personality traits, drives, emotions) and meta-reasoning shape reasoning strategies.

Today’s talks

3. Why we need Cognitive Architectures such as ACT-R for Cognitive Principles that enable dynamic & flexible Human-AI Interaction? (Nele Russwinkel)

Recent advances in generative artificial intelligence have significantly improved the fluency and responsiveness of human–AI interaction. However, purely generative approaches remain insufficient for enabling dynamic and flexible interaction in real-world environments, which are inherently uncertain, evolving, and context-dependent. Effective interaction over extended periods requires structured representations, different forms of cognitive principles, and mechanisms for adaptive control. These principles and mechanisms are relevant for good and fluent Human-AI Interaction. Humans navigate such complex uncertain environments by leveraging mental models and a sense of control, continuously adjusting their actions based on perceived affordances and situational constraints.

This talk argues that cognitive architectures, such as ACT-R (Adaptive Control of Thought—Rational), provide a critical foundation for modeling these capabilities in AI systems. By integrating symbolic and subsymbolic representations, ACT-R enables the development of systems that can maintain and update mental models, reason over context, and adapt behavior in a goal-directed manner. Such architectures support continuity, coherence, and interpretability in interaction, particularly over longer time horizons as well as anticipation of the individual in a situation. We propose that combining generative AI with cognitive architectures offers a promising path toward more robust, human-aligned AI systems capable of meaningful, sustained collaboration in dynamic environments.

4. Knowledge and reasoning across the bands of cognition (Christian Lebiere)

Human cognition involves processes and phenomena taking place at scales ranging across orders of magnitude in time and complexity that Allen Newell called the bands of cognition. In this talk, I present evidence that cognitive architectures provide a unifying framework for knowledge and reasoning across the bands of cognition. Going down to the neural band, integrating symbolic knowledge and neural-like mechanisms enables the development of neuro-symbolic architectures that combine the strengths of neural learning and generalization and symbolic representations and inference. Going up to the rational band, bounded rationality is enabled by reflecting the statistical regularities of the environment in knowledge representation and reasoning mechanisms. However, systematic deviations from rationality known as cognitive biases emerge from the interaction between knowledge and reasoning limitations of cognitive architectures. Further up into the social band, integrating large groups of interacting cognitive agents enables the emergence of social and organizational knowledge and reasoning.

Today's main questions

1. How can we explain the cognitive effects of complex information environments, especially environments containing misinformation?
2. How can computational systems perform reasoning under uncertainty?
3. How can we build generative AI systems capable of robust human-AI interaction?
4. How can cognitive architectures unify cognition across multiple scales?